COMPARATIVE EFFICIENCY OF SAMPLING PLANS (ILLUSTRATION—APPLE TREES)

ECONOMICS, STATISTICS, AND COOPERATIVES SERVICE

U.S. DEPARTMENT OF AGRICULTURE

COMPARATIVE EFFICIENCY OF SAMPLING PLANS

(ILLUSTRATION---APPLE TREES)

By

Earl. E. Houseman

Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture

September 1978

PREFACE

This publication is regarded by the author as supplementary training material for students who are familiar with, or are studying, elementary theory of sampling including stratification, cluster sampling, ratio and regression estimation, sampling with probability proportional to size, and multiple-stage sampling. After studying sampling methods one at a time, it is important to get a unified view of the several methods and the conditions under which they have about the same or different variances.

In sampling various populations we quite often find two or more techniques that are roughly equal in efficiency and reduce sampling variance about as much as possible. Administrative feasibility, costs, and freedom from potential bias are important criteria for selecting a sampling plan and become primary criteria when the choice is among plans having small differences in sampling variance.

Ability to prejudge accurately the efficiency of alternative sample designs with reference to various survey objectives and populations is important. Such ability comes from experience and detailed study of alternative techniques of sampling a population and of making estimates. Quite often only two or three alternatives are compared in an analysis because of limitations of data or only a few alternatives are of interest. In this publication many alternative

sampling and estimation plans are applied to a small population of apple trees and the results are recorded in tables for comparative purposes. The focus of attention is on the magnitude of the differences in efficiency in relation to the patterns of variation that exist.

For some readers, parts of the presentation are probably too detailed. However, it is important to understand fully the alternatives and to put mathematical expressions for estimators and their variances in forms that are most meaningful for comparative purposes. Exercises are distributed through the text.

Chapter I makes use of graphical, or geometrical, interpretations in the comparison of four alternative ways of using an auxiliary variable. There is a brief presentation of the relevant theory for each plan which is followed by a discussion of the plans including a numerical example. Sampling with probability proportional to size in comparison to other methods is of special interest. For comparison, a part of each variance formula is written as the sum of squares of deviations from a line.

Chapter II expands the comparisons made in Chapter I to include interactions in efficiency. For example, the comparative efficiency of sampling units of various size is related to the method of estimation and to stratification. Chapter III provides some further comparisons, but the emphasis is on how

theory and ingenuity solved an important problem in the sampling of fruit trees. Some comparisons involving two-stage sampling using apple trees as an example are included in Chapter IV.

This volume was written because it was a pleasure and because I always learn something from making comparisons like those contained herein.

Earl E. Houseman
Statistician

CONTENTS

CHAPTER I

SIMPLE USES OF AN AUXILIARY VARIABLE

1.1 INTRODUCTION

Proficiency in the use of auxiliary information to reduce sampling variance is an important goal in the formulation of a sampling plan. In this chapter we will compare four alternative methods of using an auxiliary variable in the design of a sample or in the estimator and one without using an auxiliary variable, giving a total of five alternative methods. The methods discussed are commonly found in textbooks on sampling. It is important to know whether an auxiliary variable is worth using and how to use it most effectively. Achievement of greater efficiency in the use of an auxiliary variable is usually inexpensive compared to increasing sample size, but incorrect use could cause an increase rather than a decrease in sampling error.

For each of the five alternatives there is an estimator and the variance of each estimator can be expressed in a form that is suitable for interpretation of the sampling variance as a function of deviations of points from a line. The emphasis in this chapter is on simple dot charts as a useful aid to understanding or judging the comparative effectiveness of alternative methods in different situations. Special attention will be given to sampling with probability proportional to size and how it compares with other ways of using an auxiliary variable

including stratification and optimum allocation. After a review of notation, definitions, and theory, a numerical example will be presented which makes use of some data collected in a research project to develop techniques for estimating apple production.

Consider a population of N sampling units and let $Y_1,\ldots, Y_N$ represent the unknown values of Y and let $X_1,\ldots, X_N$ represent the known values of an auxiliary variable X. A sample is to be selected and the values of Y for the n su's (sampling units) in the sample, namely $y_1,\ldots, y_n$, are to be obtained. The corresponding values of X for the su's in the sample are $x_1,\ldots, x_n$. We assume that the objective is to estimate the population mean, $\bar{Y} = \frac{\sum_{i}^{N} Y_i}{N}$. Also, in the interest of keeping the notation as simple as possible, let Y and X represent the population totals. That is, $Y = \sum_{i}^{N} Y_i$ and $X = \sum_{i}^{N} X_i$. This gives "Y", for example, a dual meaning as in "the characteristic Y" or as the total for the population. However, the meaning should be clear from the context.

A resume of the theory for each of the five alternatives, which will be called plans, is presented after a brief review of sampling with equal and unequal probabilities of selection.

1.1.1 EQUAL PROBABILITIES OF SELECTION

A sample obtained by selecting one su at a time, at random with equal probability and without replacement, is called a simple random sample. When the variance of Y in the

population is defined as

$$\sigma^2 = \frac{\sum_{N}(Y_i - \bar{Y})^2}{N} \qquad (1.1)$$

the variance of the mean, $\bar{y}$, of a simple random sample of n is given by

$$V(\bar{y}) = \frac{N-n}{N-1} \frac{\sigma^2}{n} \qquad (1.2)$$

If the variance of Y is defined as

$$S^2 = \frac{\sum_{N}(Y_i - \bar{Y})}{N-1} \qquad (1.3)$$

and the variance of $\bar{y}$ is

$$V(\bar{y}) = \frac{N-n}{N} \frac{S^2}{n} \quad . \qquad (1.4)$$

In the discussion that follows, S^2 will be used as the definition of the variance of y.

The mean, $\bar{y}$, of a simple random sample is an unbiased estimate of $\bar{Y}$ and the variance, $s^2 = \frac{\sum_{n}(y_i - \bar{y})^2}{n-1}$, among su's in the sample is an unbiased estimate of S^2. Incidentally, the writer from a practical point of view advises use of the word "unbiased" with some caution. In the mathematical theory, the meaning of "unbiased" is usually clear, but in practice "unbiased estimate" is often misleading to persons who are interested in estimates from a survey and are unaware of the restricted meaning of the term.[1/]

1/ See sections 4.4 and 4.5 of Expected Value of a Sample Estimate, Statistical Reporting Service, USDA, September 1974.

Exercise 1.1 Show that either definition of the variance among the N values of Y leads to the same answer for the variance of $\bar{y}$. That is, show that equations 1.2 and 1.4 are the same.

1.1.2 UNEQUAL PROBABILITIES OF SELECTION

Some sampling plans specify that sampling units be selected with pps (probability proportional to size). For simplicity, sampling with replacement is assumed.

It is often very important to make a clear distinction between the probability of selecting the i^{th} su of a population when a particular random draw is made and the probability of the i^{th} su being included in a sample. To help make the distinction clear, the letter "P" or "p" will be used to represent selection probability and "f" will represent inclusion probability, that is, the probability of any given su being in the sample. When simple random sampling is applied, each su has a probability equal to $\frac{n}{N}$ of being in the sample. That is, the inclusion probability, f, is equal to $\frac{n}{N}$ for simple random sampling.

With regard to sampling <u>with pps and replacement</u>, let P_1, $P_2,\ldots,P_N$ be the set of selection probabilities for the N su's in the population. It is specified that $\sum_{}^{N} P_i=1$. Thus, "selecting a sample with probabilities proportional to X_i" means that $P_i = \frac{X_i}{X}$ where $X=\sum_{}^{N} X_i$. Since the sampling is with replacement, the selection probabilities remain constant from one random draw to another.

The unbiased estimator of $\bar{Y}$ for a sample of n is

$$\bar{y} = \left(\frac{1}{N}\right)\left(\frac{1}{n}\right)\sum^{n}\frac{y_i}{p_i} \tag{1.5}$$

In the estimator, i is an index of the n random draws because the same su might be selected more than once. To illustrate, if on the 4th draw su number 15 in the population is selected, y_4 and p_4 are equal to Y_{15} and P_{15}. And if the 15th su is selected again on the 12th draw, y_{12} and p_{12} are equal to Y_{15} and P_{15}. In practice, techniques for avoiding the selection of the su more than once are usually introduced but such techniques are for later consideration.

Each of the n values of $\frac{y_i}{p_i}$ in Eq. 1.5 is an unbiased estimate of the population total. Thus, $\left(\frac{1}{n}\right)\sum^{n}\frac{y_i}{p_i}$ is a simple average of n independent, unbiased estimates of Y, and $\left(\frac{1}{N}\right)$ appears in Eq. 1.5 so $\bar{y}$ will be an estimator of $\bar{Y}$ instead of the population total.

The variance of $\bar{y}$, Eq. 1.5, is

$$V(\bar{y}) = \frac{\sigma^2}{n} \tag{1.6}$$

where

$$\sigma^2 = \left(\frac{1}{N^2}\right)\sum^{N}P_i\left(\frac{Y_i}{P_i} - Y\right)^2 = \left(\frac{1}{N^2}\right)\sigma_t^2 \tag{1.7}$$

and

$$Y = \sum^{N}Y_i$$

Is Eq. 1.6 reasonable? Study the estimator. For any given value of i, $\frac{y_i}{p_i}$ in repeated sampling is a random variable which has an expected value equal to the population total, Y.

By definition, the variance of $\frac{y_i}{p_i}$ is

$$\sigma_t^2 = \sum_{}^{N} P_i \left(\frac{Y_i}{P_i} - Y\right)^2$$

where i is the index to the N su's in the population. In the estimator, $\frac{1}{n} \sum_{}^{n} \frac{y_i}{p_i}$ is the average of n independent estimates; therefore, the variance of this average is $\frac{\sigma_t^2}{n}$. And, since we are interested in estimating $\bar{Y}$ rather than Y, σ_t^2 must be divided by N^2 as shown in Eq. 1.7.

1.2 RESUME OF THEORY FOR FIVE PLANS [2/]

As discussed above, we will use S^2, Eq. 1.3, as the definition of the population variance for simple random sampling with replacement and σ^2, Eq. 1.7, is the definition of population variance for sampling with pps and replacement. Notice that, when the P_i all equal $\frac{1}{N}$, σ^2 defined in 1.7 becomes 1.1.

For convenient reference, the estimators and their variances for the five plans to be discussed are listed in Table 1.1, page 29. The variances are expressed as population values (parameters) rather than as sample estimates of variance. Each variance formula is written in a form which shows a sum of squares of deviations of points from a line (or lines). Also, an alternative

---

2/ A good reference is: Cochran, W.G., Sampling Techniques: Stratified Random Sampling, Chapter 5; Ratio Estimates, Chapter 6; Regression Estimates, Chapter 7; and for sampling with probability proportional to size see Sections 9.9, 9.10, 9.11, and 9.12 of Chapter 9.

expression for the variance of the estimator for each plan is shown. For simplicity, an assumption is made that the sampling fractions are small when the sampling is without replacement. Thus, the fpc (finite population correction) factor has been omitted from the variance formulas. The fpc, namely $\frac{N-n}{N}$, can always be included if needed. Notice in Table 1.1 that, for a constant size of sample, it is only the sums of squares that differ among the plans.

A dot chart that shows one point for each pair of values of X_i and Y_i provides simple, graphical interpretations of the sums of squares in the variance formulas for the five plans. Each variance formula for the first four plans involves the deviations of Y_i from a line through the point $(\bar{X},\bar{Y})$. The fifth plan involves line segments. How do the lines for the five plans differ and how can one judge the sampling variance for one plan compared to another by looking at a dot chart?

1.2.1 PLAN 1 - MEAN ESTIMATOR

In the first three plans, simple random sampling is assumed. These three plans differ only with regard to the method of estimating $\bar{Y}$. The first plan is to use the sample average $\bar{y} = \frac{\sum_{i}^{n} y_i}{n}$ as an estimator of $\bar{Y}$. As a symbol for an estimator we will use $\hat{y}$, and a subscript will be used to distinguish the different estimators. Thus, the first estimator and its variance are

$$\hat{y}_1 = \bar{y} = \frac{\sum_{i}^{n} y_i}{n} \qquad (1.8)$$

$$V(\hat{y}_1) = \frac{S_1^2}{n} = (\frac{1}{n})\frac{\sum^{N}(Y_i-\bar{Y})^2}{N-1} \qquad (1.9)$$

The formula for the variance of $\hat{y}_1$ contains the expression $\sum^{N}(Y_i-\bar{Y})^2$. As shown in Figure 1.1, the vertical distance between a point (X_i,Y_i) and a horizontal line through $(\bar{X},\bar{Y})$ is equal to $(Y_i-\bar{Y})$. Hence $\sum^{N}(Y_i-\bar{Y})^2$ may be interpreted as the sum of squares of the deviations of Y from a horizontal line through $(\bar{X},\bar{Y})$. The closer the points are to this horizontal line, the smaller the variance of $\hat{y}_1$.

In the general context of regression estimation, Plan 1 is a special case. Cochran, in Chapter 7, Sampling Techniques, discusses regression estimation where $\hat{y}$ in the following equation is the regression estimator:

$$\hat{y} = \bar{y} + b(\bar{X}-\bar{x}) \qquad (1.10)$$

The value of the regression coefficient, b, might be preassigned or it might be computed from the sample data. It it is preassigned, b is a constant when one considers the expected value of $\hat{y}$. If b is constant, it is clear from the theory of expected values that $E(\hat{y}) = \bar{Y}$ because the expected value of $\bar{y}$ is $\bar{Y}$ and the expected value of the second term, $b(\bar{X}-\bar{x})$ is zero[3/]. Thus, the expected value of $\bar{y}$ is $\bar{Y}$ regardless of the value that is preassigned to b. There are cases where a preassigned value of b equal to 1 is of interest but that is not pertinent to the

---

3/ $E[b(\bar{X}-\bar{x})] = E(b\bar{X}) - E(b\bar{x}) = b\bar{X} - bE(\bar{x}) = 0$ because $E(\bar{x}) = \bar{X}$.

present discussion. The point of interest is that Plan 1 may be regarded as a special case of regression estimation where b is given a preassigned value equal to zero. In Plans 2 and 3, the value of b is computed from the sample.

1.2.2 PLAN 2 - RATIO ESTIMATOR

When we let b equal $\frac{\bar{y}}{\bar{x}}$, the right side of Eq. 1.10 becomes $\bar{X}\,\frac{\bar{y}}{\bar{x}}$ which is the estimator for Plan 2. Thus,

$$\hat{y}_2 = \bar{X}\,\frac{\bar{y}}{\bar{x}} \tag{1.11}$$

This estimator is called a ratio estimator since it is the ratio of two random variables $\bar{y}$ and $\bar{x}$. For simple random sampling the variance of $\hat{y}_2$ is often written as follows:

$$V(\hat{y}_2) = \frac{S_2^2}{n} = (\frac{1}{n})[S_Y^2 + R^2 S_X^2 - 2RS_{XY}] \tag{1.12}$$

where

$$S_Y^2 = \frac{\sum_{}^{N}(Y_i - \bar{Y})^2}{N-1}$$

$$S_X^2 = \frac{\sum_{}^{N}(X_i - \bar{X})^2}{N-1}$$

$$S_{XY} = \frac{\sum_{}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

and

$$R = \frac{\sum_{}^{N} Y_i}{\sum_{}^{N} X_i} = \frac{\bar{Y}}{\bar{X}}$$

The variance formula for Plan 2, Table 1.1, shows that the deviations, $(Y_i - RX_i)$, are squared and summed.

Exercise 1.2 With reference to the variance of $\hat{y}_2$, *Eq. 1.12, show that* $S_Y^2 + R^2 S_X^2 - 2RS_{XY} = \frac{\Sigma(Y_i - RX_i)^2}{N-1}$.

Consider a line through the origin and the point $(\bar{X},\bar{Y})$, see Figure 1.1. The slope of this line is $R = \frac{\bar{Y}}{\bar{X}}$. The vertical distance between this line and a point (X_i,Y_i) is $(Y_i - RX_i)$. Therefore, the sum of squares, $\Sigma(Y_i - RX_i)^2$, in the variance formula for $\hat{y}_2$ is the sum of squares of the deviations of the points (X_i,Y_i) from the line through the origin and $(\bar{X},\bar{Y})$. The only difference between the variances of $\hat{y}_1$ and $\hat{y}_2$ is the difference between $\Sigma(Y_i - \bar{Y})^2$ and $\Sigma(Y_i - RX_i)^2$. The points for the assumed population in Figure 1.1 are somewhat closer to the line through the origin and $(\bar{X},\bar{Y})$ than to a horizontal line through (X,Y). Therefore, one would expect $\hat{y}_2$ to have a smaller sampling variance than $\hat{y}_1$.

Exercise 1.3 Verify that $Y_i - RX_i$ *is the vertical distance between a point* (X_i,Y_i) *and a straight line that passes through the origin and* $(\bar{X},\bar{Y})$.

1.2.3 PLAN 3 - REGRESSION ESTIMATOR

The estimator, $\hat{y}_3$, in Plan 3 is called a regression estimator. It makes use of a line that is derived by applying the least squares method in fitting a line to the sample values of x and y. The equation for the least squares line (fitted to

the sample data) may be written as follows:

$$\hat{y}_i = \bar{y} + b(x_i - \bar{x}) \tag{1.13}$$

where $$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2}, \quad i = 1, \ldots, n$$

and $\hat{y}_i$ is the point on the line where x is equal to x_i. The estimator of $\bar{Y}$ is obtained by substituting $\bar{X}$ for x_i in 1.13 which gives

$$\hat{y}_3 = \bar{y} + b(\bar{X} - \bar{x}) \tag{1.14}$$

To understand the variance formula for $\hat{y}_3$, suppose a least squares line is determined for the population of points shown in Figure 1.1. It is

$$\hat{Y}_i = \bar{Y} + B(X_i - \bar{X}) \tag{1.15}$$

where $$B = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}, \quad i = 1, \ldots, N.$$

and $\hat{Y}_i$ is the point on the line where X is equal to X_i. This line has been determined so the sum of the squares of the deviations of Y_i from it is a minimum. That is, $\sum_{}^{N}(Y_i - \hat{Y}_i)^2$ is less than the sum of the squares of the deviations from any other straight line. The sum of squares of the deviations of Y_i from the least-squares regression line can be written as follows:

$$\sum_{}^{N}(Y_i - \hat{Y}_i)^2 = \sum_{}^{N}\{Y_i - [\bar{Y} + B(X_i - \bar{X})]\}^2 \tag{1.16}$$

The expression on the right side of 1.16 appears in the variance formula for $\hat{y}_3$ in Table 1.1 which is

$$V(\hat{y}_3) = \frac{S_3^2}{n} = (\frac{1}{n}) \frac{\sum_{N}\{Y_i - [\bar{Y}+B(X_i-\bar{X})]\}^2}{N-1} \qquad (1.17)$$

Exercise 1.4 Show that the right side of 1.16 reduces to $(1-r^2)\Sigma(Y_i-\bar{Y})^2$ *where* r *is the coefficient of correlation between* X *and* Y.

1.2.4 DISCUSSION OF PLANS 1, 2, and 3

The variances of $\hat{y}_1$, $\hat{y}_2$, and $\hat{y}_3$ have been related to the sums of squares of deviations from three lines respectively: (1) a horizontal line through $(\bar{X},\bar{Y})$, (2) a ratio line (that is, a line through the origin and $(\bar{X},\bar{Y})$, and (3) a regression line (which is a line determined by the method of least squares). Since the sum of squares of deviations from the regression line is least, the variance of $\hat{y}_3$ will generally be less than the variances for $\hat{y}_1$ and $\hat{y}_2$. The comparative variances can be judged from visual examination of how close the points are to each of the three lines.

The variance of $\hat{y}_2$ is not always less than the variance of $\hat{y}_1$. Moreover, the correlation coefficient is not a reliable measure of how the variances of $\hat{y}_1$ and $\hat{y}_2$ compare. According to Eq. 1.12, $2RS_{XY}$ must be larger than $R^2S_X^2$ or the variance of $\hat{y}_2$ will be larger than the variance of $\hat{y}_1$. In other words, use of an auxiliary variable in a ratio estimator could result in an increase rather than a decrease in variance.

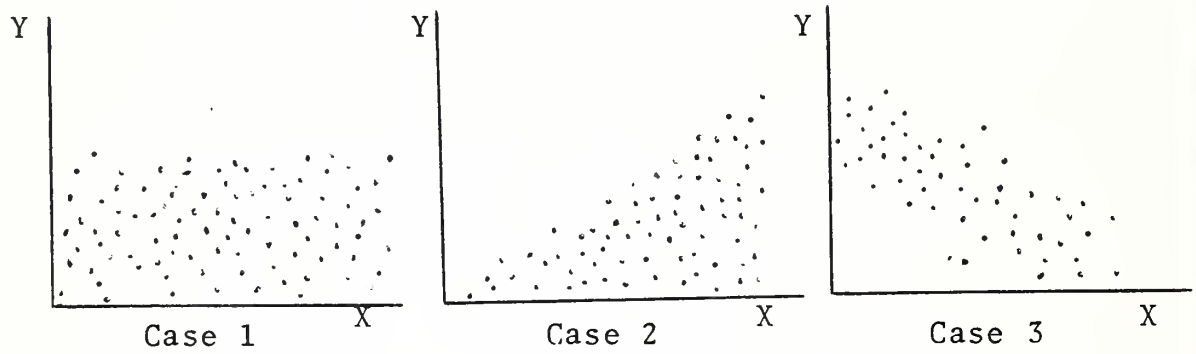
The variance formulas discussed above are population variances (parameters) which must be estimated from the sample. For all three plans, formula for estimating the sampling variances are of the same format as the population variance formula. The only difference is that the sum of squares is computed from sample data instead of data for the entire population. The variance formulas for Plans 2 and 3 are large sample approximations, which are commonly used in practice. (See Cochran's book sections 6.4 and 7.4.)

In a survey involving many variables and tabulations by various classifications, the first two estimators (plans) are commonly used. Although the variance of $\hat{y}_3$ is, to some degree, generally less than the variance of $\hat{y}_1$ or $\hat{y}_2$, its use is generally limited to special situations where low error is very important and the variance of $\hat{y}_3$ is appreciably less than the variance of $\hat{y}_1$ or $\hat{y}_2$. For example, it might be used to estimate the production of a particular commodity or when it is very important to make estimates with a high degree of accuracy for a few selected characteristics.

All three of the estimators may be used with sampling plans other than simple random sampling; for example, ratio estimators and stratified random sampling are quite common.

Exercise 1.5 For the special case where the regression line is the same as the ratio line, show that the variance of $\hat{y}_3$ *is equal to the variance of* $\hat{y}_2$. *Can* $V(\hat{y}_3)$ *ever be larger than* $V(\hat{y}_2)$*?*

Exercise 1.6 Compare Plans 1, 2, and 3 with regard to the following three dot charts representing three different relations between X *and* Y.



For each of the three cases rank the three plans from largest to smallest sampling variance.

1.2.5 PLAN 4 - SAMPLING WITH PPS

Plans 2 and 3 used the auxiliary variable in estimation and not in the design or selection of a sample. Plan 4 is to select a sample of n elements with replacement and to use probabilities of selection proportional to X_i. By substituting $\frac{x_i}{\bar{X}}$ for p_i in Eq. 1.5 and $\frac{X_i}{\bar{X}}$ for P_i in 1.7, the following expressions are obtained for the estimator and its variance:

$$\hat{y}_4 = \bar{X}\left(\frac{1}{n}\right) \sum^{n} \frac{y_i}{x_i} \tag{1.18}$$

and

$$V(\hat{y}_4) = \frac{\sigma_4^2}{n} = \left(\frac{1}{n}\right)\left(\frac{1}{N}\right) \sum^{N} \left(\frac{\bar{X}}{X_i}\right)(Y_i - RX_i)^2 \tag{1.19}$$

The formula for the variance of $\hat{y}_4$ shows that $(Y_i - RX_i)$ are the deviations which are squared. Thus, the line involved

in Plan 4 is the same as the line for the ratio estimator. Notice that the squares of the deviations, $(Y_i - RX_i)^2$, are weighted by $\frac{1}{X_i}$ owing to the unequal probability of selection. For the ratio estimator, the squares of the deviations were weighted equally. Incidentally, the appropriate formula for estimating the variance of $\hat{y}_4$ from sample data is not of the same form (and will not reduce to the same form) as Eq. 1.19.

In practice one often finds that the variance of the deviations, $(Y_i - RX_i)$, increases as X increases. That is, the values of Y are usually more widely scattered for large values of X than for small values of X. If the relation between X and Y is like the dot chart in Figure 1.2, Plan 4 will have a lower sampling variance than the first three plans. A line through $(\bar{X},\bar{Y})$ and the origin fits the data about as well as any line. But, $\hat{y}_4$ would have the least sampling variance because, as shown in the formula for its variance, the largest values of $(Y_i - RX_i)^2$ receive the smallest weights in the sum of squares. Judging the effectiveness of Plan 4 is more than a matter of observing how well the data fit a line through $(\bar{X},\bar{Y})$ and the origin. In fact, it is easy to misjudge the effectiveness of sampling with pps. We will return to this point after presentation of Plan 5.

Exercise 1.7 Start with σ^2 *as defined in 1.7 and show that it reduces to* $(\frac{1}{N})\sum^{N} \frac{\bar{X}}{X_i}(Y_i - RX_i)^2$ *when* $P_i = \frac{X_i}{X}$ *and* $R = \frac{Y}{X}$.

1.2.6 PLAN 5 - STRATIFIED SAMPLING

This plan makes use of the variable X as a basis for stratification. Suppose the sampling units in the population

have been listed in order from smallest to largest values of X. The list is then divided into L strata. Let

N_h = the population number of su's in stratum h,

n_h = the sample number of su's,

$f_h = \frac{n_h}{N_h}$ = the sampling fraction,

Y_{hi} and X_{hi} = the values of Y and X for the i^{th} su in stratum h,

S^2_{Yh} = the variance of Y within stratum h,

$\bar{Y}_h$ = the average value of Y in stratum h, and

$\bar{X}_h$ = the average value of X in stratum h.

We are primarily interested in proportional allocation of the sample to strata for comparison with Plans 1, 2, and 3, and in optimum allocation for comparison with Plan 4.

With proportional allocation the sampling fractions, f_h, are all equal and it is appropriate to use the unweighted sample mean as an estimator of $\bar{Y}$. Hence,

$$\hat{y}_5 = \bar{y} \tag{1.20}$$

Assuming simple random sampling within strata and that the fpc's are negligible,

$$V(\hat{y}_5) = \frac{S^2_5}{n} \tag{1.21}$$

where $$S^2_5 = \frac{1}{N} \Sigma N_h S^2_{Yh}$$

and $$S^2_{Yh} = \frac{\Sigma_i (Y_{hi} - \bar{Y}_h)^2}{N_h - 1}$$

With reference to a dot chart for showing deviations that are squared in the variance formulas, instead of one line, we now have a series of line segments, one for each stratum, as shown in Figure 1.3. Each line segment is a horizontal line through the stratum mean. The sampling variance, S_5^2, is an average of the squares of deviations from these horizontal line segments. If the points are close to the line segments, the sampling variance will be small for stratified random sampling.

Consider what happens to the sum of squares for stratified random sampling as the number of strata increases, that is, as the difference between the largest and smallest value of X for each stratum decreases. If the relation between X and Y over the whole population is approximately linear, the sum of squares of the deviations from the line segments will become approximately equal to the sum of squares of the deviation from a regression line as in Plan 3. Under those conditions Plans 3 and 5 would have approximately the same sampling variance. If the relation between X and Y is not linear, the sampling variance for Plan 5 might be less than the sampling variance for Plan 3, depending on the width of the stratum intervals, the degree of nonlinearity, and how close the points are to a curved line.

Suppose the ratio line (that is, a straight line through $(\bar{X},\bar{Y})$ and the origin) fits the points about as well as any line. In this case, the sampling variances for Plans 2, 3 and 5 (assuming the stratum intervals are small) would be approximately equal.

Plan 4 must be judged with regard to how well the probabilities of selection fit the situation as well as the closeness of the points to the ratio line. It is helpful to compare it with using the auxiliary variable for stratification and optimum allocation of the sample to strata. We know that the optimum size of sample from stratum h is proportional to $N_h S_{Yh}$. Or, in terms of sampling fractions, the optimum sampling fraction, f_h', is proportional to S_{Yh}.

In stratified sampling, the optimum sampling fractions are proportional to $\bar{X}_h$ when S_{Yh} is proportional to $\bar{X}_h$. In this case, the selection probabilities in sampling with pps would be approximately in proportion to $\bar{X}_h$ provided the stratum intervals are small. In other words, when S_{Yh} is proportional to $\bar{X}_h$ the optimum sampling fractions in stratified sampling are in close agreement with the selection probabilities in sampling with pps. It is very important to recognize that the situation most favorable for sampling with pps occurs when (1) the data follow the ratio line, and (2) the conditional standard deviation of Y is proportional to X. ("Conditional standard deviation" refers to the standard deviation of Y for a given value of X.) The dot chart, Figure 1.2, meets those conditions. Notice that the vertical distance between the two dotted lines is proportional to X; hence, the conditional standard deviation of Y is, at least roughly, proportional to X.

Recognition of a relation like the one in Figure 1.2 as a good case for sampling with pps provides guidance when making a

choice among alternatives including the possibility of making a transformation of X that would provide a better measure of size. Sometimes a simple transformation like $X_i' = X_i + C$, where C is a constant, will provide a measure of size, X', such that the conditional standard deviation of Y will be in proportion to X'. In some cases, a simple transformation can change sampling with pps, compared to Plan 1, from a substantial increase in sampling variance to an important reduction. With pps sampling it is important that the maximum values of Y approach zero as X approaches zero.

One might feel that sampling with probability proportional to X does not fully remove, from the sampling variance, variation among strata when X is the criterion for stratification. Look at the pps estimator. It is variation in the stratum ratios, $R_h = \frac{1}{N_h} \sum_i \frac{Y_{hi}}{X_{hi}}$ rather than variation in $\bar{Y}_h$ that needs to be considered. It will be easier to discuss this point in the next chapter when stratification in combination with different methods of estimation is considered.

Some numerical results as well as dot charts will be presented later in this chapter and in Chapter II.

Exercise 1.8 Refer to Figure 1.2 and verify from theorems pertaining to similar triangles that the range in values of Y *is proportional to* X. *In this case, as a rough approximation, we may regard the standard deviation of* Y *as being in proportion to* X. *Is it possible in sampling with probability proportional*

to X to have a lower sampling variance than sampling with stratification by X and optimum allocation? When?

Exercise 1.9 (a) Refer to Figure 1.1 and rank Plans 1, 2, 3, and 5 from least variance to highest. Ans. 3, 5, 2, 1 with 3 and 5 being close depending on the number of strata.

(b) It appears that the variance for Plan 4 would be much larger than the variance for stratified random sampling with optimum allocation. Why? Look at the conditional standard deviation of Y.

(c) Since the range in the optimum sampling fractions for stratified random sampling is small, would you agree that Plan 4 would have a much larger variance than Plan 1?

(d) Consider the simple transformation $X_i' = X_i + C$ *where C is a constant. Is there a value of C such that X' would be an effective measure of size.*

Exercise 1.10 Refer to Exercise 1.6 and for each case rank all five plans with regard to sampling variance.

Exercise 1.11 Prepare a dot chart showing a relation between X and Y such that stratified random sampling with allocation proportional to N_h, *Plan 5, will have a smaller sampling variance than the regression estimator, Plan 3.*

Exercise 1.12 Prepare a dot chart such that the variance for Plan 5 with proportional allocation will be approximately equal to the variance for Plan 1 and (at the same time) the variance for Plan 5, with optimum allocation will be much

less than the variance for Plan 1. This would be a case where gain from stratification would be entirely attributable to varying sampling fractions rather than stratification to remove variation associated with differences among stratum means.

1.2.7 SUMMARY

If there is no relation between X and Y, including a relation between X and the conditional standard deviation of Y, information about X offers no possibilities for reducing sampling variance; in fact, the sampling variance could be increased by using X. If there is a relation, some alternative ways to take advantage of it have been shown. Clearly, the most effective way of using an auxiliary variable depends on what the relation is like.

In the sampling and estimation specifications for a particular survey, an auxiliary variable would generally be used in only one way. For example, attempting to use a relationship between X and Y as a basis for stratification and also in estimation is generally not advisable. Try to fully utilize the potential contribution of an auxiliary variable in one way. Whether an auxiliary variable is used in stratification or in estimation might depend on the nature of other auxiliary variables that are available. For example, some kinds of auxiliary variables are readily useful in stratification but not estimation. Consider using quantitative measures in estimation or in sampling with pps and using nonquantitative measures in stratification. This point will receive further attention.

1.3 NUMERICAL EXAMPLE

Although our interest is in the practical application of sampling theory, a major objective in the presentation of numerical illustrations in this and later chapters is to improve one's comprehension of patterns of variation that exist and to develop one's skill at judging the effectiveness of alternative sampling and estimation methods in specific situations. It is informative to apply several alternatives to the same population even though some of the alternatives are not practically feasible.

The data for the following example were taken from a research project to develop techniques for sampling apple trees to forecast and estimate apple production. The primary purpose was to make an intensive investigation of ways of sampling a tree rather than how to select a sample of trees. As a part of this project, the branches on six apple trees were mapped. Included among the measurements that were taken are the cross-sectional area of each branch and the number of apples on each branch. There was a total of 28 primary branches on the six trees. A primary branch, which is a branch from the tree trunk, probably would not be used as a sampling unit in practice. However, data for these 28 primary branches are useful as a numerical example of alternative ways of using an auxiliary variable. Also the results will be useful in later discussions and comparisons of methods of sampling within trees.

For purposes of this numerical example, the 28 primary branches is the population of sampling units. We assume the purpose of sampling is to estimate the total number of apples on the six trees. The auxiliary variable X is the csa (cross-sectional area) of a branch. The fruit counts, Y, and the csa's, X, for the 28 limbs are presented in Table 1.2. Let us compare the five plans outlined above by referring to a dot chart. Figure 1.4 shows the points (X_i,Y_i) and three lines: (1) the horizontal line for Plan 1, (2) a ratio line through the origin and $(\bar{X},\bar{Y})$ which pertains to Plans 2 and 4, and (3) the least squares regression line for Plan 3. To order the sampling variances from smallest to largest, one would undoubtedly rank the first three plans in the order 3, 2, and 1, with 1 having a much larger variance than the other two. Since the scatter of the points increases as the csa increases, one might expect Plan 4 to be better than Plan 2, but Plan 4 is somewhat difficult to judge. In Chapter II, similar comparisons of the plans will be made using terminal branches (and hence more points) as sampling units.

The total number of sampling units, 28, is too small to provide a good example of stratified random sampling in comparison to the other four plans. However, for purposes of illustration, a comparison will be made. Since 28 is divisible by 4, it is convenient to divide the branches, after being ordered by csa, into four strata of 7 branches each as presented in Table 1.2.

The stratum boundaries are indicated by vertical dotted lines in Figure 1.4. It is evident that line segments, for the stratified random sampling as specified in the preceding paragraph, do not fit the data as well as the regression line, Plan 3. Although the sampling variance for Plan 5 is clearly much less than the variance for Plan 1, it is undoubtedly greater than the variance for Plan 3. Its rank compared to Plans 2 and 4 is uncertain.

We will now compare the judgments formed from looking at Figure 1.4 with numerical results. The relative variances of the five estimators, assuming $n = 1$ (that is, a sample of one branch), are presented in Table 1.3. Relative variances are the variances divided by $\bar{Y}^2$. Although it is not possible to select a stratified random sample of one branch, it is appropriate to let $n = 1$ for purposes of comparing Plan 5 with the other plans.

In this example, the relationship between X and Y is such that all four Plans 2, 3, 4, and 5 provide large reductions in sampling variance. Stratification, as applied, reduced the sampling variance by more than 80 percent compared to Plan 1 but not as much as Plans 2, 3, and 4 because it did not utilize as fully the information provided by X. If it were feasible to divide the population into more strata, perhaps 8 or 10 instead of 4, the relative variance for Plan 5 would have been less than 0.307 and perhaps nearly as low as the variance for the regression estimator, Plan 3. However, from the results that we have seen, it appears that the auxiliary variable X can be used to reduce

the sampling variance from 1.117 to about 0.200. Some of the practical considerations in the choice of a plan will be discussed later. In the next section our understanding of sampling with pps will be extended by comparing it to stratification with optimum allocation.

1.3.1 VARYING THE SAMPLING FRACTION WITH SIZE OF SAMPLING UNIT

From Figure 1.4 it is clear that the variance of the number of apples increases with the size of branch. The standard deviation within strata and the average csa per branch are presented in Table 1.4.

Since the largest S_{Yh} is about 10 times larger than the smallest, the largest sampling fraction (with stratified sampling and optimum allocation) would be about 10 times larger than the smallest. This range of variation in sampling fractions is large enough to expect optimum allocation, compared to proportional, to give a substantial reduction in variance. The relative variance for optimum is 0.211 compared to 0.301 for proportional.

With reference to sampling with pps, notice that the conditional standard deviation of Y is roughly in proportion to X. This is indicated by the fact that the ratio of S_{Yh} to $\bar{X}_h$, Table 1.4, is nearly constant. Also, the points in Figure 1.4 follow, approximately, a line through the origin and $(\bar{X},\bar{Y})$. Therefore, it is reasonable to find that 0.211, the variance for stratified sampling with optimum allocation, is close to 0.194, the variance for sampling with pps.

Since S_{Yh} is approximately in proportion to $\bar{X}_h$, csa is a good measure of size. However, it is informative to compare the five plans when circumference is used or a measure of size of branch. To examine the relation between number of apples and circumference, see Figure 1.5. Notice that the least squares line (Plan 3) departs farther from the origin than did the least squares line for csa, Figure 1.4. This is reflected in the variances which are presented in Table 1.5. The relative variance, 0.256, for Plan 3 is considerably less than the relative variances for Plans 2 and 4. Also notice that circumference is less effective than csa for all three Plans 2, 3, and 4.

Exercise 1.13 Refer to Table 1.2 and compute the four values of $\bar{X}_h$ *taking the circumference as the auxiliary variable. Compare these values of* $\bar{X}_h$ *with the values of* S_{Yh} *given in Table 1.4. What does this comparison indicate regarding the use of circumference as a measure of size in pps sampling?*

Notice that csa is a mathematical transformation of circumference. The question might be asked, "Is there a better transformation?" This question will be given further attention in the next chapter. For the research study, a csa measurement was made by wrapping a tape around the base of a branch. The tapes had been calibrated to give a direct reading of the csa assuming the branch is circular. Figure 1.4 suggests that csa is a good measure of size for sampling with pps, but broader experience is needed. In a later illustration it will become evident that sampling with pps is a good practical method of selecting a sample of branches.

Exercise 1.14 By careful planning one can compute sub-totals and totals of ΣY_i, ΣX_i, ΣY_i^2, $\Sigma X_i Y_i$, *and* $\Sigma \frac{Y_i^2}{X_i}$ *that provide intermediate results from which the variances for several alternative plans are easily obtained. For purposes of computation show that the values of* S^2 *for the five plans may be written as follows:*

$$S_1^2 = (\frac{1}{N-1})[\Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{N}]$$

$$S_2^2 = (\frac{1}{N-1})[\Sigma Y_i^2 - 2R\Sigma X_i Y_i + R^2 \Sigma X_i^2] \text{ where } R = \frac{\Sigma Y_i}{\Sigma X_i}$$

$$S_3^2 = S_1^2(1-r^2) \text{ where r is the correlation coefficient}$$

$$\sigma_4^2 = (\frac{\bar{X}}{N})\Sigma \frac{Y_i^2}{X_i} - \bar{Y}^2$$

Since there are 7 branches in each stratum the expression in Table 1.1 for S_5^2 *reduces to*

$$S_5^2 = (\frac{1}{N-4})[\Sigma Y_i^2 - \frac{\Sigma Y_h^2}{7}] \text{ where } Y_h \text{ is the total of Y for stratum h.}$$

From Table 1.2 the following intermediate results are obtained:

ΣY_i	= 7,199	ΣX_i	= 157.76
ΣY_i^2	= 3,844,283	ΣX_i^2	= 1,329.98
$\Sigma X_i Y_i$	= 67,633.47	$\Sigma \frac{Y_i^2}{X_i}$	= 392,247.3

The stratum totals, Y_h, *are*

$$Y_1 = 202,\ Y_2 = 923,\ Y_3 = 1{,}594,\ \textit{and}\ Y_4 = 4{,}480$$

Compute the values of S_1^2, S_2^2, S_3^2, σ_4^2, *and* S_5^2 .

Answer: $S_1^2 = 73{,}828$

$S_2^2 = 16{,}339$

$S_3^2 = 12{,}292$

$\sigma_4^2 = 12{,}826$

$S_5^2 = 20{,}274$

Table 1.1--Estimators and Their Relative Variances 1/

Plan	Estimator	Variance of estimator	S^2 or σ^2 expressed as an average of squared deviations	An alternative expression for S^2 or σ^2
1	$\hat{y}_1 = \bar{y}$	$(\frac{1}{n})S_1^2$	$S_1^2 = \frac{\Sigma(Y_i-Y)^2}{N-1}$	$S_1^2 = S_Y^2$
2	$\hat{y}_2 = \bar{X}\frac{\bar{y}}{\bar{x}}$	$(\frac{1}{n})S_2^2$	$S_2^2 = \frac{\Sigma(Y_i-RX_i)^2}{N-1}$	$S_2^2 = S_Y^2 + R^2S_X^2 - 2RS_{XY}$
3	$\hat{y}_3 = \bar{y} + b(\bar{X}-\bar{x})$	$(\frac{1}{n})S_3^2$	$S_3^2 = \frac{\Sigma\{Y_i - [\bar{Y} + B(X_i-\bar{X})]\}^2}{N-1}$	$S_3^2 = S_Y^2(1-r^2)$
4	$\hat{y}_4 = \bar{X}(\frac{1}{n})\Sigma\frac{y_i}{x_i}$	$(\frac{1}{n})\sigma_4^2$	$\sigma_4^2 = \frac{\Sigma^N(\frac{\bar{X}}{X_i})(Y_i-RX_i)^2}{N}$	$\sigma_4^2 = \frac{1}{N^2}\Sigma^N P_i(\frac{Y_i}{P_i} - Y)^2$
5	$\hat{y}_5 = \frac{\Sigma N_h\bar{y}_h}{N} = \bar{y}$	$(\frac{1}{n})S_5^2$	$S_5^2 = \frac{1}{N}\Sigma_h N_h \frac{\Sigma_i(Y_{hi}-\bar{Y}_h)^2}{N_h-1}$	$S_5^2 = \frac{1}{N}\Sigma N_h S_h^2$

1/ Upper case letters refer to population values, lower case to sample values.

N = total number of units in population.
n = total number of units in sample.
L = number of strata.
N_h = number of units in the population in stratum h.

$N = \Sigma^L N_h$

n_h = number of units in sample from stratum h.

$n = \Sigma^L n_h$

$Y = \Sigma^N Y_i$

$\bar{Y} = \frac{\Sigma Y_i}{N}$

$X = \Sigma^N X_i$

$\bar{X} = \frac{\Sigma X_i}{N}$

$P_i = \frac{X_i}{\bar{X}}$

$R = \frac{\bar{Y}}{\bar{X}}$

Y_{hi} = value of Y for the i^{th} unit in stratum h.

$\bar{Y}_h = \frac{\Sigma_i Y_{hi}}{N_h}$ = mean of Y in stratum h.

$S_Y^2 = \frac{\Sigma(Y_i-\bar{Y})^2}{N-1}$

$S_X^2 = \frac{\Sigma(X_i-\bar{X})^2}{N-1}$

$S_{XY} = \frac{\Sigma(Y_i-Y)(X_i-X)}{N-1}$

$r = \frac{S_{XY}}{(S_Y)(S_X)}$

$B = \frac{S_{XY}}{S_X^2}$

$b = \frac{s_{xy}}{s_x^2}$

$S_h^2 = \frac{\Sigma_i(Y_{hi}-\bar{Y}_h)^2}{N_h-1}$

Table 1.2--Data for Primary Branches on Six Apple Trees

(Arrayed by csa)

Stratum	Branch[1]	csa[2]	Cir.[3]	No. of apples	Stratum	Branch[1]	csa[2]	Cir.[3]	No. of apples
1	1-4	.87	3.3	5	3	6-3	4.84	7.8	183
	1-5	1.03	3.6	34		1-2	5.09	8.0	40
	5-6	1.34	4.1	4		2-4	5.75	8.5	396
	1-3	1.83	4.8	59		6-2	5.89	8.6	250
	5-4	1.83	4.8	18		2-3	6.16	8.8	157
	5-5	1.83	4.8	17		5-1	6.16	8.8	179
	6-4	1.99	5.0	65		4-2	7.18	9.5	389
2	2-5	2.68	5.8	89	4	2-2	8.94	10.6	333
	4-4	2.86	6.0	238		4-1	9.28	10.8	696
	4-5	2.86	6.0	81		2-1	9.63	11.0	473
	4-3	3.57	6.7	254		3-1	11.60	12.1	762
	1-1	3.68	6.8	76		3-3	12.84	12.7	517
	5-3	4.48	7.5	97		3-2	13.45	13.0	622
	5-2	4.72	7.7	88		6-1	15.38	13.9	1,077
					TOTAL		157.76	221.0	7,199

1/ Tree (first digit) and branch within a tree (second digit).
2/ Cross-sectional area of branch in square inches.
3/ Circumference of branch in inches.

Table 1.3--Relative Variances of Estimators

Plan	Relative variance of $\hat{y}$
1	1.117
2	0.247
3	0.186
4	0.194
5	0.307

Table 1.4--Mean csa and Standard Deviation of Y by Strata

Stratum	Mean csa, $\bar{X}_h$	Standard Deviation, S_{Yh}	$\frac{S_{Yh}}{\bar{X}_h}$
1	1.53	24.8	16.2
2	3.55	78.4	22.1
3	5.87	129	22.0
4	11.59	240	20.7

Table 1.5--Relative Variances When the Auxiliary Variable is Circumference

Plan	Relative variance of $\hat{y}$
1	1.117
2	0.559
3	0.256
4	0.438

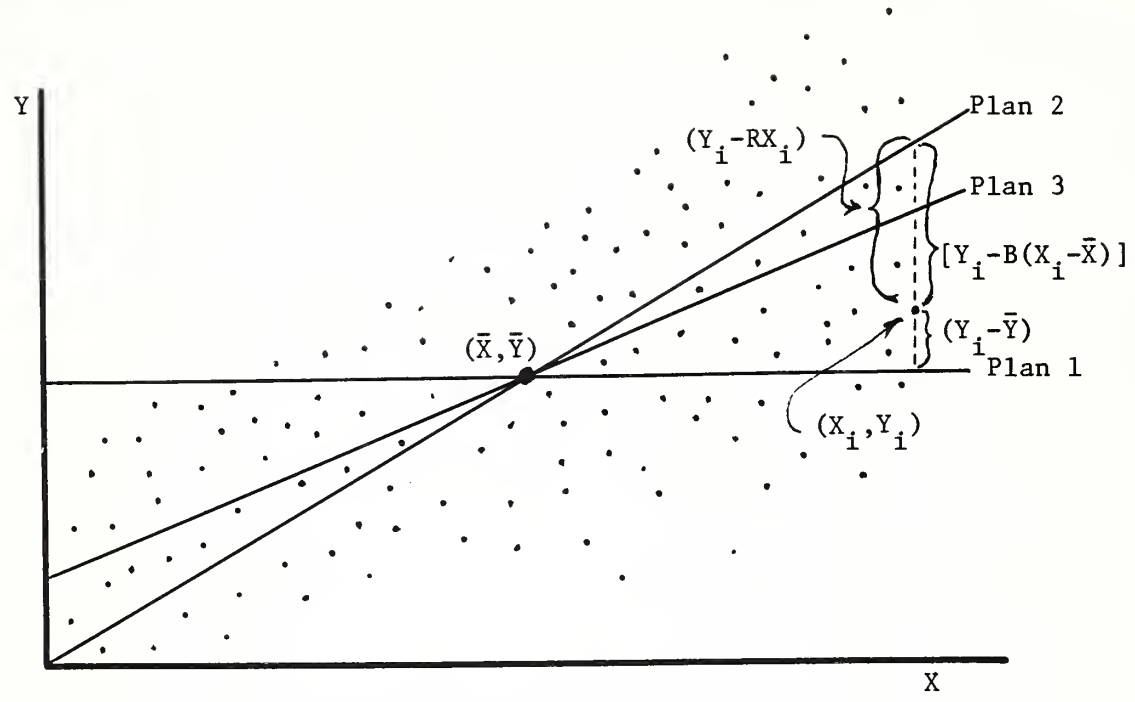


Figure 1.1--Deviations in Variance Formulas for Plans 1, 2, and 3

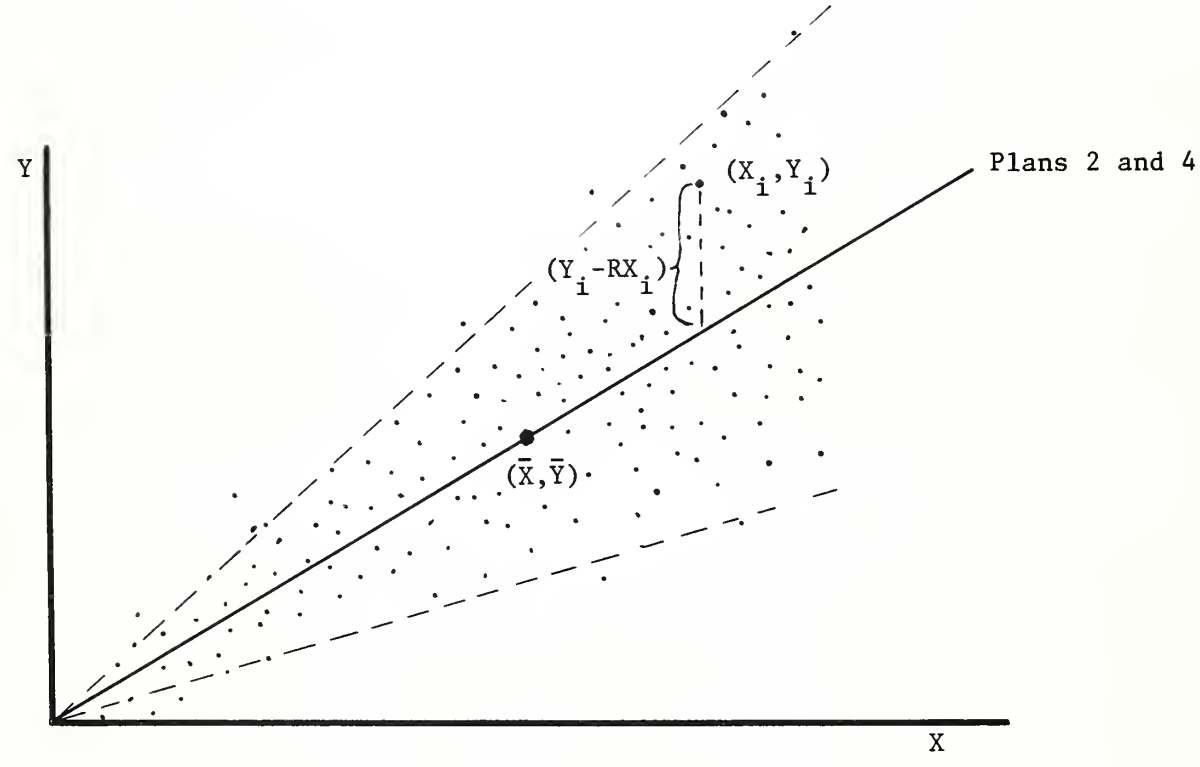


Figure 1.2--Deviations in Variance Formulas for Plans 2 and 4

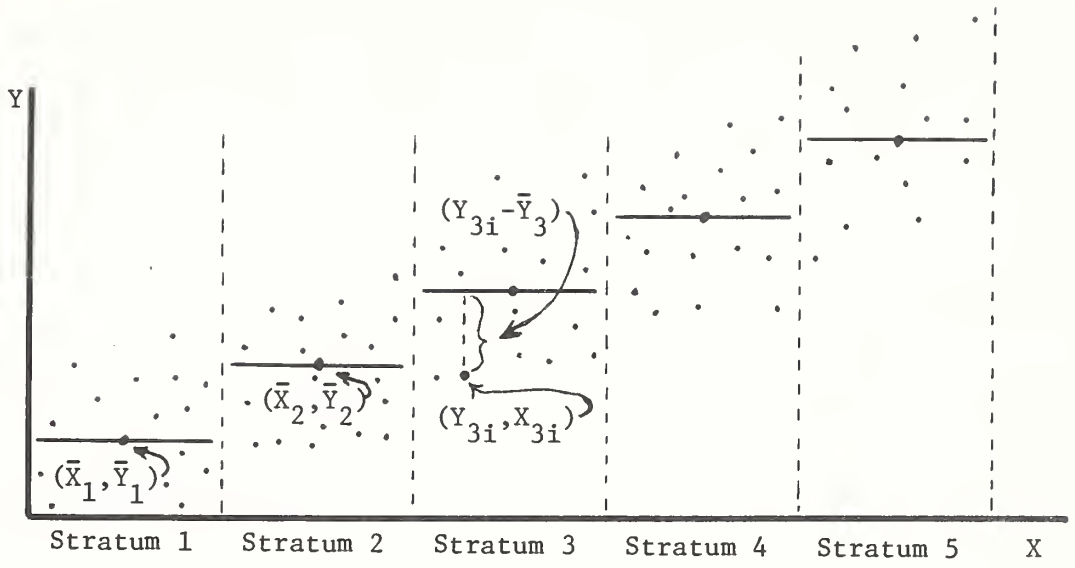


Figure 1.3--Deviations in Variance Formula
for Plan 5, Stratified Random Sampling

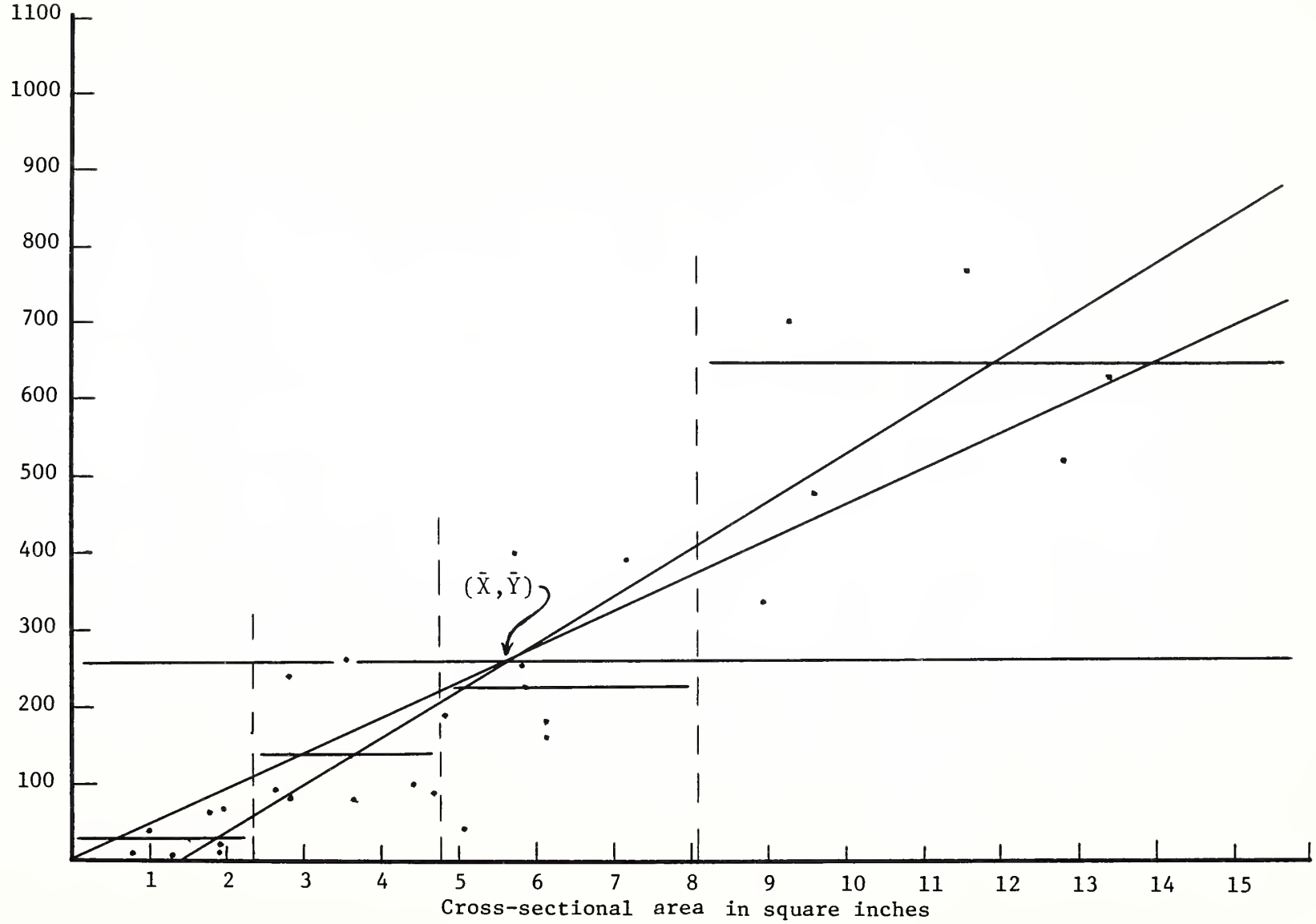


Figure 1.4--Relation between Number of Apples and CSA

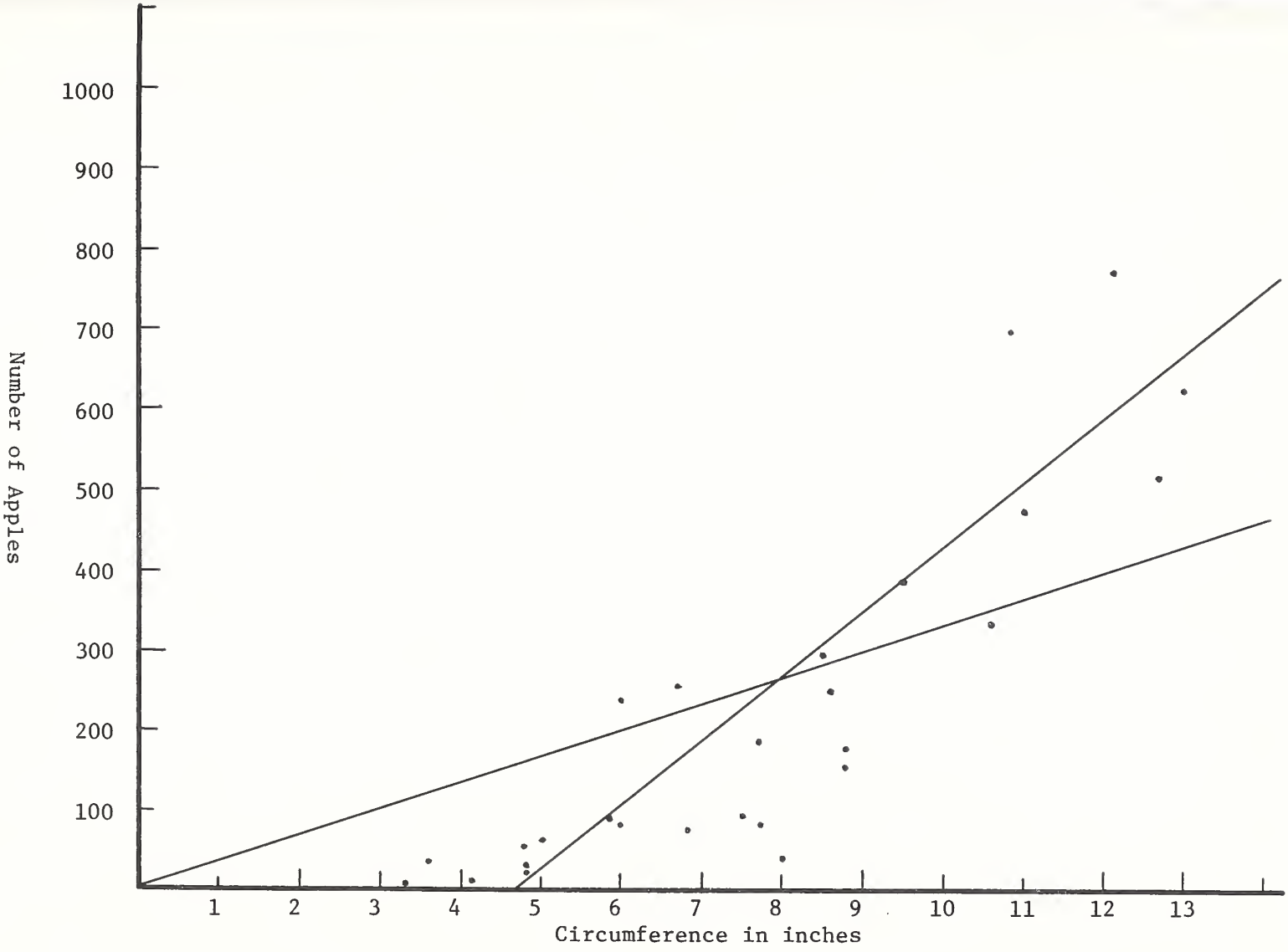


Figure 1.5--Relation between Number of Apples and Circumference

CHAPTER II

FURTHER OBSERVATIONS ON USES OF AN AUXILIARY VARIABLE

2.1 INTRODUCTION

The effects on sampling variance of various factors in sample design and estimation are not independent. For example, the difference in the sampling variance between a mean estimator and a ratio estimator might vary with the definition of the sampling unit or with the criteria used for stratification. In this chapter some numerical examples that display such interactions will be given. The objective is to further develop a perception of patterns (or components) of variation and ability to judge how alternative methods rank with regard to sampling variance. As you study and acquire experience in sampling try to visualize the pattern of variation in a population to be sampled and test your skill at prejudging the effectiveness of alternative sampling plans.

The data for the examples in this chapter are taken from the research project on methods of estimating apple production which was referred to in Chapter I. The sampling alternatives that are considered require a map of each tree that is sampled. That is, a map of a tree which defines the sampling units (branches) is the sampling frame. Methods of probability sampling are available which do not require preparing a complete map of a tree. This will be discussed in Chapter III.

As background, refer to Figure 2.1 which is a map of one of the six apple trees used for the numerical example presented in Chapter I. The map shows the scheme that was used for identifying branches. For example, 3-1-4 refers to third-stage branch number 4 from second-stage branch number 1 and first-stage branch number 3. Branches from the tree trunk were mapped until "terminal" branches were reached. "Terminal branch" refers to the last stage of branching where the mapping of branches was terminated. The csa's (cross sectional areas) of the terminal branches ranged from about 3/4 to 2 square inches which seemed to be about the smallest practical size of branch to consider as a sampling unit. There were 28 primary branches and 135 terminal branches on the six trees. The average number of apples on a terminal branch was about 50.

When following a tree trunk to primary branches, to second-stage branches, etc., small branches are sometimes found which are not large enough to be classified as terminal branches. For example, six apples were found on small branches on primary branch number 2 before the 4 second-stage branches 2-1, 2-2, 2-3, and 2-4 were reached. Apples on such branches have been called "path" fruit, meaning fruit on the path of a terminal branch. Path fruit present some special problems which will be discussed in Chapter III. The amount of path fruit is relatively small and will be ignored in this chapter.

For each of the first four plans that were discussed in Chapter I primary and terminal branches will be compared as

sampling units. Then, using terminal branches, the first four plans will then be applied within strata (trees) for comparison with each of the four plans when there is no stratification.

2.2 COMPARISON OF PRIMARY AND TERMINAL BRANCHES AS SAMPLING UNITS

The number of applies on each of the 28 primary branches and the csa of each branch were presented in Table 1.2. Data for the 135 terminal branches are presented in Table 2.1. The number of apples on primary branches included path fruit whereas the numbers on terminal branches do not. The difference is presumed to be negligible for purposes of an exercise in variance comparisons. Figures 1.4 and 2.2 are the dot charts for primary and terminal limbs respectively.

Table 2.2 presents relative variances for terminal and primary branches. The relative variances for primary branches are taken from Table 1.3 in Chapter I, and relative variances for terminal branches were computed using the same variance formulas.

When interpreting variances it is essential that the dimensions of the variances be clear. What variation does a particular variance measure and in what units is the variance expressed? Are the relative variances in Table 2.2 comparable? Let us examine the formula for the relative variance (RV) of $\hat{y}_1$, which is

$$\text{RelVar } (\hat{y}_1) = \frac{1}{\bar{Y}^2}\left(\frac{1}{n}\right)(S_1^2) \qquad (2.1)$$

where $$S_1^2 = \frac{\Sigma(Y_i - \bar{Y})^2}{N-1}$$

A quantity like S_1^2 is sometimes called "unit variance" as it is a measure of variation among individual sampling units. The quantity $\frac{S_1^2}{\bar{Y}_1^2}$ may be called "unit-relative variance" which is the square of the coefficient of variation among individual units. In Eq. 2.1, when n = 1 the relative variance of $\hat{y}_1$, is the unit-relative variance. A similar interpretation of the variance formula for the other estimators holds. Thus, S_2^2 is the unit variance that pertains to the ratio estimator, $\hat{y}_2$.

The variances presented in Table 2.2 are unit-relative variances which may be regarded as sampling variances for samples of one branch. Usually sampling variances for alternative plans are compared under one of two conditions: equal sampling fractions or equal costs. In this chapter the comparisons will be under an assumption of equal sampling fractions. The sampling fractions are $\frac{1}{28}$ and $\frac{1}{135}$, respectively, for one primary branch and one terminal branch.

To achieve comparability, the variances for primary branches will be converted to the equivalent of one terminal branch. That is, we want to find the variances for primary branches that correspond to a sampling fraction of $\frac{1}{135}$. There is an average of $\frac{135}{28} = 4.82$ terminal branches per primary branch.

Ignoring the fpc (finite population correction), the relative variance of the first estimator, $\hat{y}_1$, is $\frac{1.17}{n'}$ for a sample of n' primary branches and is $\frac{0.660}{n}$ for a sample of n terminal branches. (The numbers, 1.17 and 0.660, are from Table 2.2.) Since the sampling fractions are the same for terminal and primary branches when $n = 4.82\ n'$, we will substitute $\frac{n}{4.82}$ for n'. Thus,

$$\frac{1.17}{n'} = \frac{(4.82)(1.17)}{n} = \frac{5.639}{n} .$$

Therefore, 5.639 compares with .660 when the sampling fractions are equal. The variance, 5.639, might be described as the relative variance among primary branches expressed on the basis of one terminal branch.

The conversion factor, 4.82, also applies to the other estimators. Thus, all of the unit variances for primary branches must be multiplied by 4.82 to convert them to the equivalent of one terminal branch. This leads to Table 2.3, which reflects differences in sampling efficiency under the condition that the sampling fraction is the same for primary and terminal branches and for all four plans.

The variances in Table 2.3 are also meaningful in terms of sampling fractions that would be required when all four estimators have the same variance. Such sampling fractions would be proportional to the variances in Table 2.3, assuming the fpc's are negligible. As an example, using primary branches as sampling units, the variance of $\hat{y}_2$ will be the same as the

variance of $\hat{y}_1$ when the sampling fraction of Plan 2 is 21 percent, $\frac{1.191}{5.639} = .21$, of the sampling fraction for Plan 1. As another example, for the sampling variance of $\hat{y}_3$ using primary branches to be the same as the sampling variance of $\hat{y}_4$ using terminal branches, the sampling fraction would need to be 2.8, $\frac{0.897}{0.319}$, times larger.

Exercise 2.1 (a) Find the relative variance of $\hat{y}_4$ *for a random sample of five terminal branches. Plan 4 is sampling with pps and replacement. Ans.* 0.064.

(b) Assume simple random sampling of primary branches and find the number of primary branches so that the relative variance of $(\hat{y}_1) = 0.064$. *The answer, ignoring the fpc, is* 18.3. *There were only 28 primary branches in the population so the fpc should be taken into account. Include the fpc,* $\frac{N-n}{N}$, *in the variance formula for* $\hat{y}_1$ *and recompute the sample size that is needed. Ans.* 11 *primary branches.*

(c) With reference to (a) and (b), $135\hat{y}_4$ *and* $28\hat{y}_1$ *are estimators of the population total number of apples. Will the relative variances of these two estimators of the total be equal when the sample sizes are* 5 *terminal branches with Plan 4 and* 11 *primary branches with Plan 1?*

(d) It was stated above that, when the fpc is negligible, the variances in Table 2.3 are proportional to the sampling fractions needed to have the same relative variances of the estimates for all of the alternatives. The answers to (a) and (b) were $\frac{5}{135}$ *and* $\frac{18.3}{28}$ *when the fpc was ignored. Verify*

that these sampling fractions are proportional to the corresponding variances presented in Table 2.3.

Table 2.3 shows two major differences in efficiency: (1) Plan 1 vs the other three plans and (2) primary vs terminal branches as sampling units. Table 2.4 presents the relative variances for Plans 2, 3 and 4 as a proportion of the variances for Plan 1. Notice that the proportions of the variation among primary branches which was accounted for by variation in the size (csa) of branches was much higher than the proportions for terminal branches. Variation in the size of terminal branches was partially controlled by the specifications and process for determining a terminal branch. For primary branches the correlation between X and Y was .91. It was .69 for terminal branches

In Table 2.5 the variances for terminal branches are expressed as a proportion of the variances for primary branches. Here we see that the largest reduction in variance is under Plan 1. However, even after variance associated with variation in the csa has been taken into account in the estimator or process of selection (Plans 2, 3 and 4), the sampling variances for terminal branches are about one third of the sampling variances for primary branches. This is a manifestation of intra-class correlation--the general tendency for things that are close together in time or space to be alike. If there was no intra-class correlation, the sampling variances for Plans 2, 3 and 4 would have been about the same for primary and terminal branches. With Plan 1 the difference in variance between primary and terminal branches is attributable to the difference

in correlation between csa and number of apples as well as intra-class correlation.

The interaction shown in Table 2.3 between the variances for the four plans and the two kinds of sampling units seems typical. The situation might be viewed in this way. When the sampling units are large and auxiliary information is not used in the sample design or in estimation, the sampling variance is large and there is a large potential for reducing sampling variance. An auxiliary variable that is effective in reducing sampling variance will probably be relatively more effective when the sampling units are large. This was displayed in Table 2.4. Or, when an effective auxiliary variable is used, the relative difference in sampling variance between large and small sampling units will probably be less as displayed in Table 2.5.

The same phenomenon has been observed in various other situations. In area sampling, for example, if geographic stratification is effective, it will tend to be relatively more effective when the area sampling units are large than when they are small. This is not a justification for large sampling units. The implication is that matters of sample design and estimation are more critical when the sampling units are large and vary widely in size.

There is a limit to the reduction in variance that can be achieved through sample design and estimation techniques. That is, assuming a fixed sampling fraction, one might imagine

a practical minimum variance as a goal to be achieved by design. There might be a number of alternatives which will approach that goal. Table 2.3 shows three alternatives with relative variances between 0.3 and 0.4.

Exercise 2.2 Table 2.6, which will be discussed later, shows the number of apples on each of the six trees in the column headed Y_h. *The variances among trees,* $\frac{\Sigma(Y_i-\bar{Y})^2}{N-1}$, *is* 464,295, *where* Y_i *is the number of apples on the* i^{th} *tree and* N *is the number of trees. Verify that the relative variance of* Y_i *is* 0.344. *This is the relative variance of* $\hat{y}_1$ *when a tree is the sampling unit and the size of the sample is one tree. Convert this variance,* 0.344, *to the equivalent of one terminal branch. Ans.* 7.74. *Compare the answer with the variances in Table 2.3 for Plan 1.*

Exercise 2.3 Assume that a simple random sample of terminal branches on the six trees is to be selected and that $N\bar{y}$ *is the estimator of the total number of apples on the six trees. Ignoring the fpc, how many terminal branches need to be selected so the variance of* $N\bar{y}$ *is equal to the variance of an estimate based on a random selection of one tree and a count of all apples on the tree? Assume that* 6y *is the estimator for the sample of one tree where* y *is the number of apples on the sample tree. Refer to exercise 2.2 for the variance among trees and to Table 2.2 for the variance among terminal branches. Ans. The variance of an estimate from a sample of* 2 *terminal branches is equal to the variance of an estimate from a sample*

of one tree. There were 22.5 *terminal branches per tree so* 2 *terminal branches is less than one-tenth of one tree. This result is typical of the low sampling efficiency of a large sampling unit. Moreover, it is very difficult to make an accurate count of all apples on a tree.*

2.3 STRATIFICATION BY TREES

Table 2.6 presents variances, covariances, and other information for each of the six trees. These data pertain to terminal branches. They will be used to determine the variances for five different estimators based on stratified random sampling with trees as strata and a constant sampling fraction. Stratified sampling with pps within trees will also be considered which gives a total of six alternatives. For these six alternatives, designated as plans 6 through 11, we want to find sampling variances that are comparable with the variances presented in Table 2.2 for nonstratified sampling of terminal branches.

It is advantageous to become sufficiently familiar with sampling theory to avoid searching textbooks for a formula and checking it to be sure it is applicable. A formula as found in a textbook might be appropriate but need adaptation. By recalling a few things from the theory of random variables, correct variance formulas can be readily derived for finding the sampling variances for the sampling and estimation plans that follow.

For comparative purposes, relative variances are recorded in Table 2.7 for four plans that have been discussed and for six additional plans that will be discussed in the next section.

2.3.1 PLAN 6--MEAN ESTIMATOR

In Plan 6 the sample is allocated to trees (strata) in proportion to the number of terminal branches on the trees. You may notice that Plan 6 is the same as Plan 5 except that the strata are trees instead of size-of-branch classes. The estimator of the population mean, $\bar{Y}$, is

$$\hat{y}_6 = \sum_h \frac{N_h}{N} \bar{y}_h = \frac{N_1}{N} \bar{y}_1 + \ldots + \frac{N_L}{N} \bar{y}_L \qquad (2.2)$$

where $\bar{y}_h = \frac{\sum_i y_{hi}}{n_h}$ is the sample average for stratum h (i.e., the average number of apples per terminal branch on tree h),

h is the index for strata (trees),

i is the index for sampling units within stratum h (branches on tree),

N_h is the total number of sampling units (terminal branches) in stratum h,

$N = \Sigma N_h$ is the total number of sampling units in the population, and

n_h is the number of sampling units in the sample from stratum h (number of branches in the sample from tree h).

Exercise 2.4 Since the sample is stratified and allocated to strata in proportion to N_h, *the estimator is a simple average of all values of* y_{hi} *in the sample. Show that this is true.*

The estimator, $\hat{y}_6$, was written as shown in Eq. (2.2) because, to find its variance, we need to consider it as a function of the stratum means. The weights $\frac{N_h}{N}$, are constant. Therefore, the variance of $\hat{y}_6$ depends on the variance of the stratum means. The sample from one stratum is independent of the sample from another stratum. Therefore, the stratum means, $\bar{y}_h$, are independent random variables, and the terms, $\frac{N_h}{N}\bar{y}_h$, in $\hat{y}_6$ are independent random variables. We know from the theory of random variables that the variance of the sum of independent random variables is the sum of the variances of the random variables. This gives the basis for writing the variance of $\hat{y}_6$ as follows:

$$V(\hat{y}_6) = \sum_h [V(\frac{N_h}{N}\bar{y}_h)] = V(\frac{N_1}{N}\bar{y}_1) + \ldots + V(\frac{N_L}{N}\bar{y}_L)$$

We also know that the variance of a constant times a variable equals the square of the constant times the variance of the variable. Hence,

$$V(\hat{y}_6) = \sum_h [(\frac{N_h}{N})^2 V(\bar{y}_h)] = (\frac{N_1}{N})^2 V(\bar{y}_1) + \ldots + (\frac{N_L}{N})^2 V(y_L) \qquad (2.3)$$

Next, we need an expression for the variance of $\bar{y}_h$. Since the sample within each stratum is a simple random sample, the variance of $\bar{y}_h$ is as follows:

$$V(\bar{y}_h) = \left(\frac{N_h - n_h}{N_h}\right)\frac{S_{Yh}^2}{n_h} \qquad (2.4)$$

where $$S_{Yh}^2 = \frac{\sum_i (Y_{hi} - \bar{Y}_h)^2}{N_h - 1}$$

The subscript Y in S_{Yh}^2 is included to show that the variance refers to the variable Y. Later we will need to take the variance of X into account and will use S_{Xh}^2 to represent the variance of X within stratum h and S_{XYh} to represent the covariance of X and Y within stratum h.

For simplicity and convenience assume that the sampling fractions, $f_h = \frac{n_h}{N_h}$, are small so the fpc's, $\frac{N_h - n_h}{N_h}$, may be ignored. Thus, dropping the fpc and substituting the variance of $\bar{y}_h$ in Eq. (2.3) gives:

$$V(\hat{y}_6) = \Sigma\left(\frac{N_h}{N}\right)^2 \frac{S_{Yh}^2}{n_h} = \left(\frac{N_1}{N}\right)^2 \frac{S_{Y1}^2}{n_1} + \ldots + \left(\frac{N_L}{N}\right)^2 \frac{S_{YL}^2}{n_L} \qquad (2.5)$$

Since the sampling specifications called for a constant sampling fraction, $\frac{n_h}{N_h}$ is constant from stratum to stratum which means that

$$\frac{n_1}{N_1} = \ldots = \frac{n_L}{N_L} = \frac{n}{N}$$

where $$\Sigma n_h = n \quad \text{and} \quad \Sigma N_h = N$$

Substituting $\frac{n}{N}$ for $\frac{n_h}{N_h}$ in Eq. (2.5) and simplifying the expression we obtain

$$V(\hat{y}_6) = \frac{1}{n}\,\Sigma\,\frac{N_h}{N}\,S_{Yh}^2 = \frac{1}{n}\left[\frac{N_1}{N}\,S_{Y1}^2 + \ldots + \frac{N_L}{N}\,S_{YL}^2\right] \qquad (2.6)$$

Exercise 2.5 Perform the algebra that is necessary to go from Eq. (2.5) to Eq. (2.6).

For Plan 6, let

$$S_6^2 = \Sigma W_h S_{Yh}^2 = W_1 S_{Y1}^2 + \ldots + W_L S_{YL}^2 \qquad (2.7)$$

where $W_h = \frac{N_h}{N}$

Since S_{Yh}^2 will be replaced by corresponding variances that are involved later in Plans 7 and 8, let $S_{6h}^2 = S_{Yh}^2$ so the notation will reflect the number of the plan or estimator. Then Eq. (2.7) becomes

$$S_6^2 = \Sigma W_h S_{6h}^2 \qquad (2.8)$$

and Eq. (2.6) simplifies to the following form

$$V(\bar{y}_6) = (\frac{1}{n}) S_6^2 \qquad (2.9)$$

where S_6^2 is a weighted average of the within stratum variances. The values of S_{6h}^2 are recorded in Table 2.6 in the column headed S_{6h}^2 and the value of S_6^2 is 1367 which is recorded in the line labeled "Separate." The reason for calling this line "Separate" will be explained later.

For purposes of comparing variances for alternative plans the choice of a sample size is arbitrary. Previously, the sampling variances for alternative plans were compared assuming n = 1. Even though it is impossible to select a stratified random sample of only one unit, it is possible to let n = 1 in Eq. 2.9 and regard the variance of $\hat{y}_6$ as the sampling

variance for a hypothetical sample of one unit. As with simple random sampling, a stratified random sample of n units would have a sampling variance equal to $\frac{1}{n}$ times the sampling variance for a hypothetical stratified random sample of one unit--provided n is large enough so the n_h, which must be integers, are approximately in proportion to N_h. Remember, these numerical examples are being worked as though the sampling fraction, f_h, is constant and small.

Exercise 2.6 Calculate the variance of $\hat{y}_6$ *assuming* n = 1. *In other words, find the value of* S_6^2. *Also, calculate the relative variance of* $\hat{y}_6$ *when* n = 1. *Your answer should agree with the relative variance of* $\hat{y}_6$ *which is recorded in Table 2.7.*

Exercise 2.7 Since $\hat{y}_6$ *is an estimate of* $\bar{Y}$, $N\hat{y}_6$ *is an estimate of the population total. Find the standard error of* $N\hat{y}_6$ *for* n = 1. *Ans.* 4991.

2.3.2 PLAN 7--RATIO ESTIMATORS BY STRATA

Plan 7 is the same as Plan 6 except that $(\bar{X}_h \frac{\bar{y}_h}{\bar{x}_h})$, instead of $\bar{y}_h$, is used in Eq. 2.2 as an estimator of the stratum mean, $\bar{Y}_h$. Thus,

$$\hat{y}_7 = \Sigma \frac{N_h}{N}(\bar{X}_h \frac{\bar{y}_h}{\bar{x}_h}) = \frac{N_1}{N}(\bar{X}_1 \frac{\bar{y}_1}{\bar{x}_1}) + \ldots + \frac{N_L}{N}(X_L \frac{\bar{y}_L}{\bar{x}_L}) \qquad (2.10)$$

The derivation of the relative variance of $\hat{y}_7$ follows the derivation in Plan 6. Simply replace the variance of $\bar{y}_h$ in

Eq. (2.3) with the variance of $(\bar{X}_h \frac{\bar{y}_h}{\bar{x}_h})$. The variance of $(\bar{X}_h \frac{\bar{y}_h}{\bar{x}_h})$, ignoring the fpc, is

$$V(\bar{X}_h \frac{\bar{y}_h}{\bar{x}_h}) = (\frac{1}{n_h}) S_{7h}^2$$

where $$S_{7h}^2 = S_{Yh}^2 + R_h^2 S_{Xh}^2 - 2R_h S_{XYh}$$

Notice that S_{7h}^2 is the same as S_2^2 in Table 1.1 except that S_{7h}^2 is a variance within stratum h rather than a variance over the whole population. Substituting S_{7h}^2 for S_{Yh}^2 in Eqs. (2.5), (2.6), and (2.7) leads to the following results:

$$V(\hat{y}_7) = (\frac{1}{n}) S_7^2 \qquad (2.11)$$

where $$S_7^2 = \Sigma W_h S_{7h}^2$$

The values of S_{7h}^2 and S_7^2 are presented in Table 2.6.

Exercise 2.8 The estimator $\hat{y}_7$, Eq. (2.10), was expressed in a form to show its similarity to $\hat{y}_6$. Is there a modification of Eq. (2.10) that would be better for computing the value of $\hat{y}_7$ from sample data? How would you compute the value of $\hat{y}_7$ from a sample?

Exercise 2.9 From the data presented in Table 2.6, find the relative variance of $\hat{y}_7$ for n = 1. *The answer,* 0.279, *is in Table 2.7. How would you explain why the sampling variance for $\hat{y}_7$ is less than the sampling variance for $\hat{y}_6$?*

The sampling variance for Plan 2 (no stratification and the ratio estimator) was 0.382 compared with 0.279 for Plan 7 (stratification and separate ratio estimators by strata). The geometrical interpretation of the sum of squares for Plan 7 compared with Plan 6 is analogous to Plan 2 compared with Plan 1. Horizontal lines for Plan 6 (one for each tree) are replaced by lines through the origin and the stratum means of X and Y. With the ratio estimator, $\hat{y}_7$, the effect of stratification depends on how much the ratio lines differ among strata. More will be said later about stratification and ratio estimators.

Exercise 2.10 Notice with reference to Eq. (2.10) that $N_h\bar{X}_h$ *is the population total of* X *for stratum* h. *Let* $X_h = N_h\bar{X}_h$ *and substitute* X_h *in Eq. (2.10) which gives*

$$\hat{y}_7 = \frac{1}{N} \Sigma X_h \frac{\bar{y}_h}{\bar{x}_h} = \frac{1}{N} \left[X_1 \frac{\bar{y}_1}{\bar{x}_1} + \ldots + X_L \frac{\bar{y}_L}{\bar{x}_L}\right]$$

With $\hat{y}_7$ *in this form, write a formula for the variance of* $\hat{y}_7$.

2.3.3 PLAN 8--REGRESSION ESTIMATORS BY STRATA

Plan 8 is like Plans 6 and 7 except that the regression estimator (see Plan 3, Chapter I) is used stratum by stratum. Thus, instead of Eq. (2.2) or (2.10) we have

$$\hat{y}_8 = \Sigma\left\{\frac{N_h}{N} [\bar{y}_h + b_h(\bar{X}_h - \bar{x}_h)]\right\}$$

$$= \frac{N_1}{N} [\bar{y}_1 + b_1(\bar{X}_1 - \bar{x}_1)] + \ldots + \frac{N_L}{N} [\bar{y}_L + b_L(\bar{X}_L - \bar{x}_L)] \qquad (2.12)$$

In the derivation of the variance of $\hat{y}_8$, the variance of $\bar{y}_h+b_h(\bar{X}_h-\bar{x}_h)$ replaces the variance of $\bar{y}_h$ in Eq. (2.3). This leads to an equation for the variance of $\hat{y}_8$ which is similar to the variances of $\hat{y}_6$ and $\hat{y}_7$. Thus,

$$V(\hat{y}_8) = \frac{1}{n} S_8^2 \qquad (2.13)$$

where $S_8^2 = \Sigma W_h S_{8h}^2$

and $S_{8h}^2 = S_{Yh}^2(1-r_h^2)$

where r_h is the correlation between X and Y within stratum h.

Exercise 2.11 Find the relative variance of $\hat{y}_8$ *for* n = 1. *Compare your result with the relative variance for* $\hat{y}_8$ *that is recorded in Table 2.7.*

2.3.4 DISCUSSION OF PLANS 6, 7, and 8

Compare the estimators, $\hat{y}_6$, $\hat{y}_7$, and $\hat{y}_8$, and their variances with $\hat{y}_1$, $\hat{y}_2$, and $\hat{y}_3$, and their variances, Table 2.7. In essence each stratum in Plans 6, 7, and 8 is treated as a separate population and the estimators and their variances within each of the strata are combined using appropriate weights. Geometric interpretations of the sampling variances with reference to sums of squares is analogous to the interpretations given in Chapter I for Plans 1, 2, and 3. There is one line for each stratum and each of the estimators, $\hat{y}_6$, $\hat{y}_7$, and $\hat{y}_8$.

Figure 2.3 presents a dot chart for each of the six trees. For each tree, the solid line is the ratio line involved in $\hat{y}_7$ and the broken line is the ratio line for $\hat{y}_2$, no stratification. As recorded in Table 2.7, the relative variances of $\hat{y}_7$ and $\hat{y}_2$ are 0.279 and 0.382 respectively. This indicates the degree to which the 6 ratio lines fit the data better than the single line. Figures analogous to Fig. 2.3 could be prepared for $\hat{y}_6$ compared with $\hat{y}_1$, for $\hat{y}_8$ compared with $\hat{y}_3$, for $\hat{y}_8$ compared with $\hat{y}_7$, etc.

Separate stratum estimators like $\hat{y}_7$ and $\hat{y}_8$ are seldom used in practice. However, Plans 7 and 8 were included for comparative purposes and further understanding of possible alternatives. There will be additional discussion of these plans after Plans 9, 10, and 11 have been presented.

2.3.5 PLAN 9--COMBINED RATIO ESTIMATOR

Instead of making a ratio estimate for each stratum and combining the separate stratum estimates, the data from the strata are combined before computing a ratio. Likewise, in Plan 10, results for individual strata are combined and used to determine a "combined regression estimator." This explains the two titles "Separate" and "Combined" in Table 2.6. The "Separate" line contains averages of within stratum variances for Plans 7, 8, and 11 which use separate stratum estimators. The entries in the "Combined" line pertain to the combined stratum estimators in Plans 9 and 10. The distinction between separate and combined is not applicable to the mean estimator, Plan 6.

s_6^2 is shown in both lines of the table.

The "combined ratio estimator" is

$$\hat{y}_9 = \bar{X} \frac{\bar{y}_s}{\bar{x}_s} \tag{2.14}$$

where

$$\bar{y}_s = \Sigma \frac{N_h}{N} \bar{y}_h$$

$$\bar{x}_s = \Sigma \frac{N_h}{N} \bar{x}_h$$

The letter "s" in $\bar{y}_s$ and $\bar{x}_s$ is used to indicate that $\bar{y}_s$ and $\bar{x}_s$ are means that pertain to a stratified random sample.

To find the variance of $\hat{y}_9$, it is convenient to remember that the large sample approximation of the relative variance (RelVar) of the ratio of any two random variables u and v is

$$\text{RelVar}(\frac{u}{v}) = \text{RelVar}(u) + \text{RelVar}(v) - 2\text{RelCov}(u,v)$$

Therefore, since $\bar{y}_s$ and $\bar{x}_s$ are random variables we have

$$\text{RelVar}(\frac{\bar{y}_s}{\bar{x}_s}) = \text{RelVar}(\bar{y}_s) + \text{RelVar}(\bar{x}_s) - 2\text{RelCov}(\bar{y}_s, \bar{x}_s) \tag{2.15}$$

Exercise 2.12 Verify that the relative variance of $\hat{y}_9$ *is equal to the relative variance of the ratio,* $\frac{\bar{y}_s}{\bar{x}_s}$.

With reference to Eq. 2.15, notice that $\bar{y}_s$ is the same as $\hat{y}_6$. We found for Plan 6, Exercise 2.6, that the RelVar of $\hat{y}_6$, and therefore of $\bar{y}_s$, was 0.512 for n = 1. The RelVar of $\bar{x}_s$ is determined in the same way. According to Table 2.6, the average

within stratum variance of X is 0.2566 and $\bar{X}^2 = 2.0240$. Thus, RelVar($\bar{x}_s$) for n = 1 is $\frac{0.2566}{2.0240} = 0.127$.

Exercise 2.13 Find the average within stratum covariance of X and Y in Table 2.6. Then find the value of RelCov($\bar{y}_s, \bar{x}_s$) *for* n = 1. *Ans.* 0.166.

According to Eq. (2.15) the RelVar of $\frac{\bar{y}_s}{\bar{x}_s}$ is

$$0.512 + 0.127 - 2(0.166) = 0.307$$

Therefore, the RelVar of $\hat{y}_9$ is 0.307. This answer is recorded in Table 2.7.

Exercise 2.14 Start with Eq. 2.15 and show that the variance of $\hat{y}_9$ *is given by*

$$V(\hat{y}_9) = V(\bar{y}_s) + R^2 V(\bar{x}_s) - 2R[Cov(\bar{y}_s, \bar{x}_s)]$$

where $R = \frac{\bar{Y}}{\bar{X}}$. *Suggestion: Notice that* $V(\hat{y}_9) = \bar{Y}^2[RelVar(\hat{y}_9)]$, *then multiply the right hand side of Eq. (2.15) by* $\bar{Y}^2$. *For* n = 1 *the value of* $V(\bar{y}_s)$, $V(\bar{x}_s)$, $Cov(\bar{y}_s, \bar{x}_s)$ *and* R *are given in the "Combined" line of Table 2.6. Using these values compute the value of* $V(\hat{y}_9)$. *The answer is* 817, *which is also in the "Combined" line.*

Exercise 2.15 Beginning with the variance of $\hat{y}_9$ *as expressed algebraically in Exercise 2.14, show that*

$$V(\hat{y}_9) = \frac{1}{n} S_9^2$$

where $$S_9^2 = \sum_h W_h S_{9h}^2$$

and $$S_{9h}^2 = S_{Yh}^2 + R^2 S_{Xh}^2 - 2RS_{XYh}$$

Suggestion: Since $\bar{y}_s$ *and* $\hat{y}_6$ *are the same you may refer to Eq. (2.9) to obtain the appropriate formula for* $V(\bar{y}_s)$. *Formulas for* $V(\bar{x}_s)$ *and* $Cov(\bar{x}_s,\bar{y}_s)$ *are analogous. Compare* S_{9h}^2 *with* S_{7h}^2.

Exercise 2.16 Continuing from the formula for $V(\hat{y}_9)$ *which is given in Exercise 2.15, show that*

$$V(\hat{y}_9) = \frac{1}{n}\left[\sum_h W_h \frac{\sum_i (Y_{hi} - RX_{hi})^2}{N_h - 1}\right]$$

The formula for the variance of $\hat{y}_9$, which is given in Exercise 2.16, shows that the deviations which are squared are deviations from a line through the origin and $(\bar{X},\bar{Y})$ where $\bar{X}$ and $\bar{Y}$ are the overall means of X and Y. This line for the combined ratio estimator, $\hat{y}_9$, is the same as the line pertaining to $\hat{y}_2$, the ratio estimator without stratification. Thus, if $\hat{y}_9$ has a lower variance than $\hat{y}_2$ it is attributable to the effect of stratification which assures proportional representation in the sample by strata. That is, there is proportional representation by strata of the deviations of Y_i from the combined ratio line. In Plan 7 there was proportional representation and separate ratio lines by strata. RelVar of $\hat{y}_9$ was 0.307 compared to 0.382 for $\hat{y}_2$ and 0.279 for $\hat{y}_7$ (see Table 2.7).

2.3.6 PLAN 10--COMBINED REGRESSION ESTIMATOR

As in the case of the combined ratio estimator, data for strata may be combined and a single (or combined) regression used instead of separate regressions. The estimator, $\hat{y}_{10}$, for the combined regression looks like $\hat{y}_3$ but it is an average within stratum regression that is determined from combined within stratum variances and covariances. Since the sampling fraction is constant, the appropriate weights for combining the within stratum variances and covariances are $\frac{N_1}{N}, \ldots, \frac{N_L}{N}$ which are the same weights used previously for combining variances. The combined within stratum variances of Y and X, 1367 and .2566, and the combined within stratum covariance, 12.23, are shown in the "Combined" line of Table 2.6. These numbers are needed for computing the "Combined" regression coefficient, the "Combined" correlation coefficient, and the variance of $\hat{y}_{10}$ for n = 1, which are also shown in the "Combined" line of Table 2.6. The corresponding numbers for sampling without stratification are shown in the last line of Table 2.6.

Algebraically

$$\hat{y}_{10} = \bar{y}_s + b_s(\bar{X}-\bar{x}_s) \qquad (2.16)$$

where

$$b_s = \frac{\bar{S}_{XY}}{\bar{S}_X^2}$$

$$\bar{S}_{XY} = \sum_h W_h S_{XYh}$$

$$\bar{S}_X^2 = \sum_h W_h S_{Xh}^2 \quad \text{and}$$

$$W_h = \frac{N_h}{N}$$

Notice that lower case letters, x and y, are used in the definition of b_s to indicate that it is computed from sample values. In Table 2.6, the value of B_s is shown which is the population value that b_s is an estimate of. The bar in the expression $\bar{S}_{XY}$ and $\bar{S}_X^2$ indicates that $\bar{S}_{XY}$ and $\bar{S}_X^2$ are averages of within stratum covariances and variances. (Previously, we had used S_Y^2, S_X^2, and S_{XY} to represent the overall variances and covariances without stratification.) The subscript "s" is used as a code indicating that stratified random sampling and combined-stratum estimation are involved. To recapitulate, b_h is a least squares estimate of the regression coefficient within stratum h, b_s is an estimate of the combined regression coefficient in the combined regression estimator, and b is the least-squares regression coefficient computed from a simple random sample without stratification.

The variance of $\hat{y}_{10}$ is

$$V(\hat{y}_{10}) = (\frac{1}{N})S_{10}^2 \qquad (2.17)$$

where

$$S_{10}^2 = \bar{S}_Y^2 (1-r_s^2)$$

$$\bar{S}_Y^2 = \sum_h W_h S_{Yh}^2$$

and

$$r_s = \frac{\bar{S}_{XY}}{\sqrt{\bar{S}_X^2 \bar{S}_Y^2}}$$

The variance of $\hat{y}_{10}$ involves squares of the deviations of Y_i from a line with a slope equal to B_s that passes through $(\bar{X},\bar{Y})$.

Remember the assumptions and that the variance formula is a large sample approximation. Further discussion of the basis for the formula for the variance of $\hat{y}_{10}$ will be omitted. For more detail the reader is referred to Cochran.[4/]

Exercise 2.17 Verify that the regression coefficient for the combined within stratum regression is 47.7 *and that the combined within stratum correlation coefficient is* 0.653. *Then verify that the relative variance of* $\hat{y}_{10}$ *is* 0.294 *for* n = 1.

2.3.7 PLAN 11--SAMPLING WITH PPS WITHIN STRATA

As in Plan 4, sampling with replacement is assumed for simplicity. That is, in stratum h the probability of the i^{th} sampling unit being selected on any given draw is proportional to X_{hi}. The estimator is

$$\hat{y}_{11} = \sum_h \left(\frac{N_h}{N}\right) \hat{y}_{11h} \qquad (2.18)$$

where $$\hat{y}_{11h} = \left(\frac{\bar{X}_h}{n_h}\right) \sum_i \frac{y_{hi}}{x_{hi}}$$

Notice that $\hat{y}_{11h}$ is like $\hat{y}_4$, the difference being that $\hat{y}_{11h}$ is an estimate of the stratum mean $\bar{Y}_h$ whereas $\hat{y}_4$ is an estimate of the population mean $\bar{Y}$. Also, notice that the estimator, $\hat{y}_{11}$, can be obtained by substituting $\hat{y}_{11h}$ for $\bar{y}_h$ in Eq. 2.2.

4/ Cochran, W. G., Sampling Techniques, Second Edition, Chapter 7. John Wiley & Sons, Inc., 1963.

It follows that the variance of $\hat{y}_{11}$ can be obtained by substituting the variance of $\hat{y}_{11h}$ for the variance of $\hat{y}_h$ in Eq. (2.3).

Owing to the similarity of $\hat{y}_{11h}$ and $\hat{y}_4$ the formula for the variance of $\hat{y}_4$ is applicable. Simply add a subscript h in the formula for the variance of $\hat{y}_4$, which gives

$$V(\hat{y}_{11h}) = (\frac{1}{n_h})(\frac{1}{N_h^2})\Sigma_i P_{hi}(\frac{Y_{hi}}{P_{hi}} - Y_h)^2$$

where $$P_{hi} = \frac{X_{hi}}{X_h}$$

$$Y_h = \Sigma_i Y_{hi}$$

and $$X_h = \Sigma_i X_{hi}$$

Substitution of $V(\hat{y}_{11h})$ in Eq. (2.3) gives

$$V(\hat{y}_{11}) = \Sigma_h (\frac{N_h}{N})^2 [(\frac{1}{n_h})(\frac{1}{N_h^2})\Sigma_i P_{hi}(\frac{Y_{hi}}{P_{hi}} - Y_h)^2] \qquad (2.19)$$

As in the derivation of 2.6, assume that n_h is proportional to N_h, which means that $n_h = (\frac{n}{N})N_h$. Substituting $(\frac{n}{N})N_h$ for n_h in Eq. (2.19) leads to the following which expresses the variance of $\hat{y}_{11}$ in a form like that used for the other estimators:

$$V(\hat{y}_{11}) = (\frac{1}{n})S_{11}^2 \qquad (2.20)$$

where $$S_{11}^2 = \Sigma_i W_h S_{11h}^2$$

and $$S_{11h}^2 = \frac{1}{N_h^2}\Sigma_i P_{hi}(\frac{Y_{hi}}{P_{hi}} - Y_h)^2$$

Like the variances for the other estimators, S^2_{11} is a weighted average of the appropriate within stratum variances. The within stratum variances, S^2_{11h}, are presented in Table 2.6 and S^2_{11}, the variance of $\hat{y}_{11}$ for n = 1, is recorded in the "separate" line.

Exercise 2.18 Using the data in Table 2.1, find the value of S^2_{11h} *for* h = 1. *Check your answer with the value recorded in Table 2.6. Is there a better expression than the one given above for finding the values of* S^2_{11h}*?*

Exercise 2.19 From the data by individual trees that are presented in Table 2.6, find the RelVar of $\hat{y}_{11}$ *for* n = 1 *and check your answer with the value of* S^2_{11} *that is recorded in Table 2.7.*

A geometrical interpretation of the variance of $\hat{y}_{11}$ is a matter of making an interpretation for each stratum and judging the average situation over all strata. For this purpose reference is made to the discussion of $\hat{y}_4$ in Chapter I.

2.3.8 SUMMARY AND DISCUSSION

Sampling variances for 10 out of 11 plans are presented in Table 2.7 for terminal branches as sampling units. Plan 5 was not applied to terminal branches. All of the plans have an important practical shortcoming. It is necessary to define, label, and list all terminal branches on a tree before it is sampled. Some ways of avoiding this will be discussed in Chapter III. However, Chapter II was intended as an exercise

in the use of theory to find the variances for alternative sampling and estimation plans and as a study of the differences in the variances for several alternatives.

As you gain experience through evaluations of sampling plans you will become increasingly aware of prevailing patterns of variation. You will observe manifestations of the general tendency for things to be stratified in space or time, or the tendency for things that are close together in space or time to be alike. There are exceptions. For example, in a field where the plant population is very dense there might be a negative intra-plot correlation among plants within very small plots owing to competition between adjacent plants.

From the results in Table 2.7 we find that the two plans with the largest variance are Plans 1 and 6. Neither plan makes use of size of branch as an auxiliary variable. The reductions in variance from use of csa as an auxiliary variable are substantial. This strongly suggests exploration of practical ways of using csa as a measure of branch size unless it is possible and feasible, when determining terminal branches, to restrict the sizes within narrow limits.

The variances for "separate" ratio and regression estimators are moderately less than the corresponding variances for the "combined" ratio and regression estimators. A small difference in favor of "separate" estimators is indicated by general experience and mathematical considerations. However, "combined" ratio or regression estimators are generally used

in practice because: (1) they are more convenient; (2) in some situations, bias in the "separate" estimators is appreciable relative to the standard error; and (3) the variance formulas, which are large sample approximations, are better approximations for the "combined" estimators. Separate ratio or regression estimators might be preferable to combined estimators when the number of strata is very small and the ratios or regressions differ widely among the strata.5/

With stratified random sampling, sampling variance is a function of variation within strata. It is generally better to judge the impact of stratification by considering within stratum variation than by the differences among strata. Making a choice between two alternative methods of stratification solely on the basis of differences among strata could in some cases be misleading. For example, the variance among the means of four strata could be much larger than the variance among the means of 30 strata. That does not necessarily mean that the sampling variance for the four strata will be the least. Also, the effect of stratification and of optimum allocation among strata are not independent of the method of estimation.

In the preceding discussion, stratification was considered as a matter of reducing sampling variance. In the design of a sample for a survey, attention needs to be given to the domains (subpopulations) for which estimates are important.

5/ See Cochran, Sampling Techniques, for a discussion of the properties of the separate and combined estimators.

This might be a primary determiner of the stratification and allocation of the sample. The sampling variances of domain estimates depend, among other things, on how close the boundaries of strata for sampling purposes correspond to the domains for which estimates are required.

2.4 FURTHER COMPARISON OF SAMPLING WITH PPS TO STRATIFIED SAMPLING WITH OPTIMUM ALLOCATION

In Chapter 1, sampling with PPS was compared to stratified random sampling with optimum allocation to strata. The data for terminal branches provide a better set of data for study of sampling with pps including the possibility of a transformation of X or Y to reduce sampling variance. For this purpose, five size-of-branch strata based on csa will be used. Table 2.8 shows the definition of these strata and presents key information about the five strata. (Reference is made to Sections 1.2.5, 1.2.6, and 1.3.1 on sampling with pps and stratified random sampling with optimum allocation.) For stratified random sampling using the mean estimator, Eq. 2.2, the optimum sampling fraction for stratum h is given by

$$f_h' = \left(\frac{S_{Yh}}{\Sigma N_h S_{Yh}}\right) n$$

In previous comparisons we assumed n = 1. Hence, we are interested in the values of

$$f_h' = \frac{S_{Yh}}{\Sigma N_h S_{Yh}} \qquad (2.21)$$

which are presented in Table 2.9. The values of f_h' are the sampling fractions within strata for a hypothetical sample of one branch and are comparable to the probabilities $P_i = \frac{X_i}{X}$ for selecting one branch with pps, where X_i is the csa of the i^{th} branch and $X = \sum_{i}^{N} X_i$. Let P_h equal the average value of P_i for the branches in stratum h. The values of P_h are presented in Table 2.9 for comparison with values of f_h'.

Exercise 2.20 Verify the values of f_h' *and* P_h *in Table 2.9 for one or two of the strata. The data presented in Table 2.8 are sufficient for this purpose.*

The values of f_h' and P_h agree quite well, which means the probability of a branch being in a sample is roughly the same for both methods, except for variation of P_i within a stratum. The next question is how well do the lines that are involved fit the data. Turn to Figure 2.4. The points appear to fit horizontal lines (which are not shown) for stratified random sampling approximately as well as the line through the origin and $(\bar{X},\bar{Y})$ for pps.

The variance for sampling with pps, Plan 4, has already been obtained. According to Table 2.2 it is 0.319. For the stratified sampling with optimum allocation, which will be called Plan 12, the estimator of $\bar{Y}$ is

$$\hat{y}_{12} = (\frac{1}{N})\Sigma N_h \bar{y}_h \qquad (2.22)$$

and the variance of $\hat{y}_{12}$ is given by

$$V(\hat{y}_{12}) = \frac{1}{n}(S_{12}^2) \qquad (2.23)$$

where $$S_{12}^2 = (\frac{\Sigma N_h \ S_{Yh}}{N})^2$$

Exercise 2.21 Using the data in Table 2.8, find the RelVar of $\hat{y}_{12}$ *for* n = 1. *Ans.* 0.273.

Exercise 2.22 From Eqs. (2.5) and (2.21) derive algebraically the variance of $\hat{y}_{12}$ *which is given by 2.23.*

In this example, the RelVar for stratified random sampling with optimum allocation was 0.273 compared with 0.319 for sampling with pps. The difference in variance is attributable to the difference in probabilities of selection and to how well the lines that are implicitly involved fit the data.

A question posed earlier was whether some transformation of X might provide a better measure of size for pps sampling. A simple transformation would be $X_i' = X_i + C$, where C is a constant and X_i' is the transformed variable. The least squares regression line, see Figure 2.4, crosses the horizontal axis at 0.46. Since the least squares line fits the data "better" than any other straight line the transformation X_i - 0.46 is suggested. With this transformation the least squares and pps lines become the same. But, with the transformation X_i - 0.46 the maximum values of Y_i do not approach zero as $X_i' = X_i$ - 0.46 approaches zero which indicates that the transformation is not a good one. However, it is informative to

make the transformation. Instead of $P_i' = \frac{X_i'}{\bar{X}}$, we now have $P_i' = \frac{X_i' - 0.46}{\Sigma(X_i - 0.46)}$. The average values of P_i' within strata, which are labeled P_h', are presented in Table 2.9 for comparison with the values of P_h' and the optimum sampling fractions, f_h'.

When X_i is used as the auxiliary variable, the relative variance is 1.403 compared to 0.319 when X_i is the auxiliary variable. The relative variance for Plan 1 was only 0.660. Thus, the transformed variable X_i' results in an increase in variance compared to simple random sampling with equal probabilities of selection. Before transformation, the selection probabilities for sampling with pps were .0032 and .0140 for the smallest and largest branches and were .0012 and .0171 after transformation. Thus, the range in the selection probabilities were greatly increased by the transformation from a factor of $4.3 = \frac{.0140}{.0032}$ to a factor of $14.2 = \frac{.0171}{.0012}$. The transformation does not effect the optimum sampling fractions, f_h'.

Exercise 2.23 Verify two of the values of P_h' *that are presented in Table 2.9.*

Exercise 2.24 Verify that when the transformation $X_i' = X_i - 0.46$ *is made the pps line and the least squares regression line become the same.*

With reference to Figure 2.4, consider two lines through the origin that represent maximum and minimum values of Y. (See Figure 1.2 in Chapter I which was portrayed as a good case for pps). As an approximation, theory suggests that a good measure of size is one that is proportional to the difference between two lines representing the maximum and minimum values of Y. A look at Figure 2.4 with this in mind suggests that a transformation such as $X_i' = X_i - 0.46$ is not a good possibility for reducing variance.

Exercise 2.25 Refer to Table 2.8 and for each stratum divide S_{Yh} *by* $\bar{X}_h$ *and examine* S_{Yh} *as a proportion of* $\bar{X}_h$. *Does this indicate that a transformation of* X *would be advisable? Ans.* No. *A transformation of* X *is not indicated and one should accept csa as a measure of size unless there is evidence to the contrary from other sources.*

Table 2.10 provides a comparison of variances for four plans when X and $X' = X - 0.46$ are the auxiliary variables.

Exercise 2.26 Explain why the transformation of X *to* X′ *has no effect on the variances for Plans 3 and 12.*

Exercise 2.27 From the data presented in the last line of Table 2.6, compute the RelVar of the ratio estimator (Plan 2) after the transformation, that is, when $X_i' = X_i - 0.46$ *is used as an auxiliary variable instead of* X_i. *Notice that a transformation of this kind does not affect the variance or covariance, but the value of* R *is changed.*

We have established that the csa of a branch is a good measure of size with regard to probabilities of selection. Judging from Figure 2.4, the least squares line and the ratio line differ enough to raise a question of transforming Y instead of X. It would be possible to add a constant to Y so the least squares regression line for X and Y′ (where Y′= Y+C) would pass through the origin and be the same as the ratio line involving X and Y′. Such a transformation would not change the selection probabilities and it appears that an appreciable reduction in variance might be obtained.

Exercise 2.28 Find the value of C *in* Y′= Y+C *such that the regression line for* X *and* Y′ *will pass through the origin. Ans.* C = 24.

Consider the result from Exercise 2.28 and the transformation Y′= Y + 24. The estimator of $\bar{Y}$, without stratification, would be

$$\hat{y} = \frac{1}{n} (\bar{X}) \left(\sum_{i}^{n} \frac{y_i + 24}{x_i} \right) - 24$$

The RelVar of $\hat{y}$ is 0.294 which is about 8 percent less than 0.319, the RelVar for Plan 4. If it is necessary to estimate C from the sample, the variance of C must be taken into account and there is no gain from the transformation. However, if there is good prior information about the value of C, the possibility of the transformation Y+C might be worth considering.

When considering sampling with pps, look for a measure of size that is close to being proportional to Y_i. If Y_i is

exactly proportional to X_i, the ratio $\frac{Y_i}{X_i}$ is constant and the sampling variance is zero. Also, the situation is a good one for sampling with pps when the standard deviation of Y for a fixed value of X is proportional to the value of X.

In stratified random sampling some statisticians, in the absence of a better basis, allocate a sample to strata according to estimates of the proportion of the total that each stratum accounts for. For example, stratum 5 accounts for 23 percent, $\frac{1634}{6973}$, of the apples so 23 percent of the sample would be allocated to stratum 5. Only 8 percent of the sample would be allocated to stratum 1 even though it contains 21 percent of the branches. In practice, prior data often provide an estimate of the proportion of the population total that each stratum accounts for. Thus, the size of sample, n_h'', for stratum h would be

$$n_h'' = P_h'' n \tag{2.24}$$

where P_h'' is the proportion that stratum h accounts for and n is the total size of the sample. The sampling fraction for stratum h is

$$f_h'' = (\frac{n_h''}{N_h}) = (\frac{P_h''}{N_h}) n$$

and for n = 1

$$f_h'' = \frac{P_h''}{N_h}$$

Thus, $f_h'' = \frac{P_h''}{N_h}$ compares with the sampling fractions (or selection probabilities) for the methods discussed previously. Values of f_h'' for the numerical example on apples are in Table 2.9.

Exercise 2.29 Verify two of the values of f_h''.

Allocating a sample to strata in proportion to prior estimates of the amounts that the strata account for is a good plan where the coefficients of variations of Y are nearly constant among strata. In Table 2.8 notice that the coefficient of variations tend to decrease as the branch size increases. This phenomena appears very frequently. The fact that the first stratum has the highest coefficient of variation means that the sample for the first stratum should be larger than n_h'', given by Eq. 2.24. This is also indicated by the comparison of f_h' and f_h''. As a "rule of thumb," some statisticians have adopted a practice of allocating a sample according to (2.24) and then doubling the size of the sample for the first stratum and increasing the sample for the second stratum by 50 percent. Small departures from an optimum allocation have a negligible impact on variance. Moreover, in practice an exact optimum allocation cannot be achieved because exact values of within stratum variances are unknown.

Table 2.1--Number of Apples and Cross Section Areas
(Terminal Branches)

Tree No. 1		Tree No. 2		Tree No. 3		Tree No. 4		Tree No. 5		Tree No. 6	
CSA	No. of Apples	CSA	No. of Apples	CSA	No. of Apples	CSA	No. of Apples	CSA	No. of Apples	CSA	No. of Apples
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
.62	2	1.83	67	2.68	206	1.93	97	1.20	30	1.09	42
.97	15	1.99	45	.97	32	2.41	165	2.15	41	.87	25
1.27	12	1.12	51	1.48	73	2.24	124	.97	10	1.47	56
.72	32	1.93	54	2.32	138	2.41	143	1.15	8	1.91	116
.62	14	3.26	116	1.83	133	2.07	58	1.27	40	1.09	27
.97	5	1.14	51	.97	32	1.54	75	1.15	14	1.15	46
.87	8	1.34	80	1.03	30	1.47	59	1.15	36	.62	2
.72	9	.87	0	1.43	27	1.43	92	.87	0	2.24	112
1.01	11	.92	58	2.24	88	1.47	82	.72	15	1.76	83
.97	7	1.76	19	.92	42	1.29	58	.92	2	.87	36
1.83	59	1.15	45	1.99	109	1.09	30	.67	13	.67	35
.87	5	1.17	33	1.47	74	1.21	36	1.54	57	1.03	59
1.03	34	1.03	41	1.64	56	.76	24	1.21	31	1.22	29
		.62	42	1.54	116	2.41	230	1.61	8	1.00	25
		1.14	12	1.99	124	.72	35	1.54	32	1.09	61
		1.27	25	1.83	79	1.03	75	1.54	26	1.09	27
		2.41	54	1.47	30	1.91	128	1.83	18	1.54	59
		1.27	14	1.21	31	1.18	0	1.83	17	1.34	25
		1.35	35	1.91	41	1.16	47	1.34	4	1.68	77
		1.07	39	1.47	16	1.15	34			2.68	120
		1.03	30	1.15	23					1.40	59
		1.91	28	1.40	35					2.07	64
		2.07	129	1.42	61					1.47	83
		1.67	108	1.76	116					1.21	19
		1.61	123	3.26	88					1.83	16
		1.61	63	2.59	50					1.03	26
		1.27	26							.97	50
										.62	36
										1.27	48
										1.99	65
$N_1 = 13$		$N_2 = 27$		$N_3 = 26$		$N_4 = 20$		$N_5 = 19$		$N_6 = 30$	

$$S_Y^2 = \frac{\Sigma(Y_i - \bar{Y})2}{N-1} = 1762$$

$$S_X^2 = \frac{\Sigma(X_i - \bar{X})2}{N-1} = .2890$$

$$S_{XY} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{N-1} = 15.46$$

$N = 135$

$\Sigma Y = 6973$

$\Sigma X = 192.06$

$R = 36.31$

$r = .6850$

$B = 53.49$

Table 2.2--Relative Variances of Numbers of Apples Among Primary and Terminal Branches

Plan	Estimator	Relative Variances Among	
		Primary Branches	Terminal Branches
1	$\hat{y}_1$	1.17	0.660
2	$\hat{y}_2$	0.247	0.382
3	$\hat{y}_3$	0.186	0.350
4	$\hat{y}_4$	0.194	0.319

Table 2.3--Relative Variances of Numbers of Apples Expressed in Terms of One Terminal Branch

Plan	Estimator	Relative Variances Among	
		Primary Branches	Terminal Branches
1	$\hat{y}_1$	5.639	0.660
2	$\hat{y}_2$	1.191	0.382
3	$\hat{y}_3$	0.897	0.350
4	$\hat{y}_4$	0.935	0.319

Table 2.4--Relative Variances Expressed as a Proportion of the Variances for Plan 1

Plan	Primary Branches	Terminal Branches
1	1.00	1.00
2	0.21	0.58
3	0.16	0.53
4	0.17	0.48

Table 2.5--Relative Variances for Terminal Branches as a Proportion of the Variances for Primary Branches

Plan	Primary Branches	Terminal Branches
1	1.00	0.12
2	1.00	0.32
3	1.00	0.39
4	1.00	0.34

Table 2.6--Data for Six Trees [1/]

Tree	N_h	Y_h	X_h	S^2_{6h} (Also S^2_{Yh})	S^2_{Xh}	S_{XYh}	R_h	S^2_{7h}	r_h	S^2_{8h}	B_h	S^2_{11h}
1	13	213	12.47	259	0.1019	3.50	17.08	169	0.681	139	34.3	139
2	27	1388	39.81	1147	0.3089	10.10	34.88	818	0.537	816	32.7	772
3	26	1850	43.97	2184	0.3301	16.64	42.07	1368	0.620	1344	50.4	1120
4	20	1592	30.88	3106	0.2996	25.67	51.55	1256	0.842	904	85.7	1038
5	19	402	24.66	241	0.1563	2.66	16.30	196	0.433	196	17.0	182
6	30	1528	40.27	892	0.2479	11.22	37.95	397	0.755	384	45.3	364
Separate [2/]	N			S^2_6				S^2_7		S^2_8		S^2_{11}
	135	xxxx	xxxxx	1367	xxxxxx	xxxxx	xxxxx	745	xxxxx	682	xxxx	644
Combined [3/]	N	Y	X	S^2_6	$\bar{S}^2_X$	$\bar{S}_{XY}$	R	S^2_9	R_s	S^2_{10}	B_s	
	135	6973	192.06	1367	0.2566	12.33	36.31	817	0.653	784	47.7	xxxx
Overall [4/]	N	Y	X	S^2_1	S^2_X	S_{XY}	R	S^2_2	r	S^2_3	B	S^2_4
	135	6973	192.06	1762	0.2890	15.46	36.31	1020	0.685	935	53.49	852

1/ See text for explanation of symbols.

2/ Results in this line pertain to separate stratum estimators. S^2_6, S^2_7, S^2_8, are averages of the corresponding within stratum variances using $\frac{N_h}{N}$ as weights. See Equations 2.6 and 2.7 and Exercise 2.13.

3/ Results in this line pertain to combined stratum estimators, $\hat{y}_9$ and $\hat{y}_{10}$.

4/ No stratification. Results pertain to the first four plans.

Table 2.7--Summary of Sampling Plans--Estimators and Their Relative Variances for a Sample of One Terminal Branch

Plan	Method of Sampling 1/	Estimator 2/		Relative Variance for n = 1
1	A	mean	$\hat{y}_1 = \bar{y}$	0.660
2	A	ratio	$\hat{y}_2 = \bar{X} \frac{\bar{y}}{\bar{x}}$	0.382
3	A	regression	$\hat{y}_3 = \bar{y} + b\ (\bar{X} - \bar{x})$	0.350
4	B	p.p.s	$\hat{y}_4 = \bar{X}(\frac{1}{n})\ \Sigma\frac{y_i}{x_i}$	0.319
6	C	mean	$\hat{y}_6 = \bar{y} = \frac{1}{N}\ \Sigma N_h \bar{y}_h$	0.512
7	C	separate ratios	$\hat{y}_7 = \frac{1}{N}\Sigma X_h \frac{\bar{y}_h}{\bar{x}_h}$	0.279
8	C	separate regressions	$\hat{y}_8 = \Sigma\ \frac{N_h}{N}\ [\bar{y}_h + b_h(\bar{X}_h - x_h)]$	0.256
9	C	combined ratio	$\hat{y}_9 = \bar{X} \frac{\bar{y}_s}{\bar{x}_s}$	0.307
10	C	combined regression	$\hat{y}_{10} = \bar{y}_s + b_s\ (\bar{X} - \bar{x}_s)$	0.294
11	D	p.p.s	$\hat{y}_{11} = \Sigma_h\ \frac{N_h}{N}\ (\frac{\bar{X}_h}{n_h})\Sigma\frac{y_i}{x_i}$	0.241

1/ A Simple random sampling without stratification
B Sampling with replacement and p.p.s without stratification
C Simple random sampling within strata (trees)
D Sampling with p.p.s within strata (trees)

2/ h is the index to strata
s as in $\bar{y}_s$ refers to stratification. Thus $\bar{y}_s$ is the mean of a stratified random sample and $\bar{y}$ is the mean of a simple random sample.

Table 2.8--Data for Size-of-Branch Strata

Stratum	Branch Size	No. of Branches	$\bar{X}_h$	No. of Apples	$\bar{Y}_h$	S_{Yh}	$\frac{S_{Yh}}{\bar{Y}_h}$
1	.60-1.00	28	0.816	586	20.93	16.79	0.80
2	1.01-1.40	45	1.161	1500	33.33	17.22	0.52
3	1.41-1.80	28	1.543	1765	63.04	31.42	0.50
4	1.81-2.20	21	1.935	1488	70.86	39.53	0.56
5	2.21+	13	2.550	1634	125.69	52.72	0.42
Total or Average		135	1.423	6973	51.65		

Table 2.9--Sampling Fractions

Stratum	f'_h	P_h	P'_h	f''_h
1	.0046	.0042	.0027	.0030
2	.0047	.0060	.0054	.0048
3	.0086	.0080	.0083	.0090
4	.0109	.0101	.0113	.0102
5	.0145	.0133	.0161	.0180

Table 2.10--Effect of Transformation

Estimator	Plan	Relative Sampling Variance	
		Before Transformation	After Transformation
Ratio(no stratification)	2	0.382	0.350
Regression(no stratification)	3	0.350	0.350
PPS (no stratification)	4	0.319	1.403
Mean (stratification by size)	12	0.273	0.273

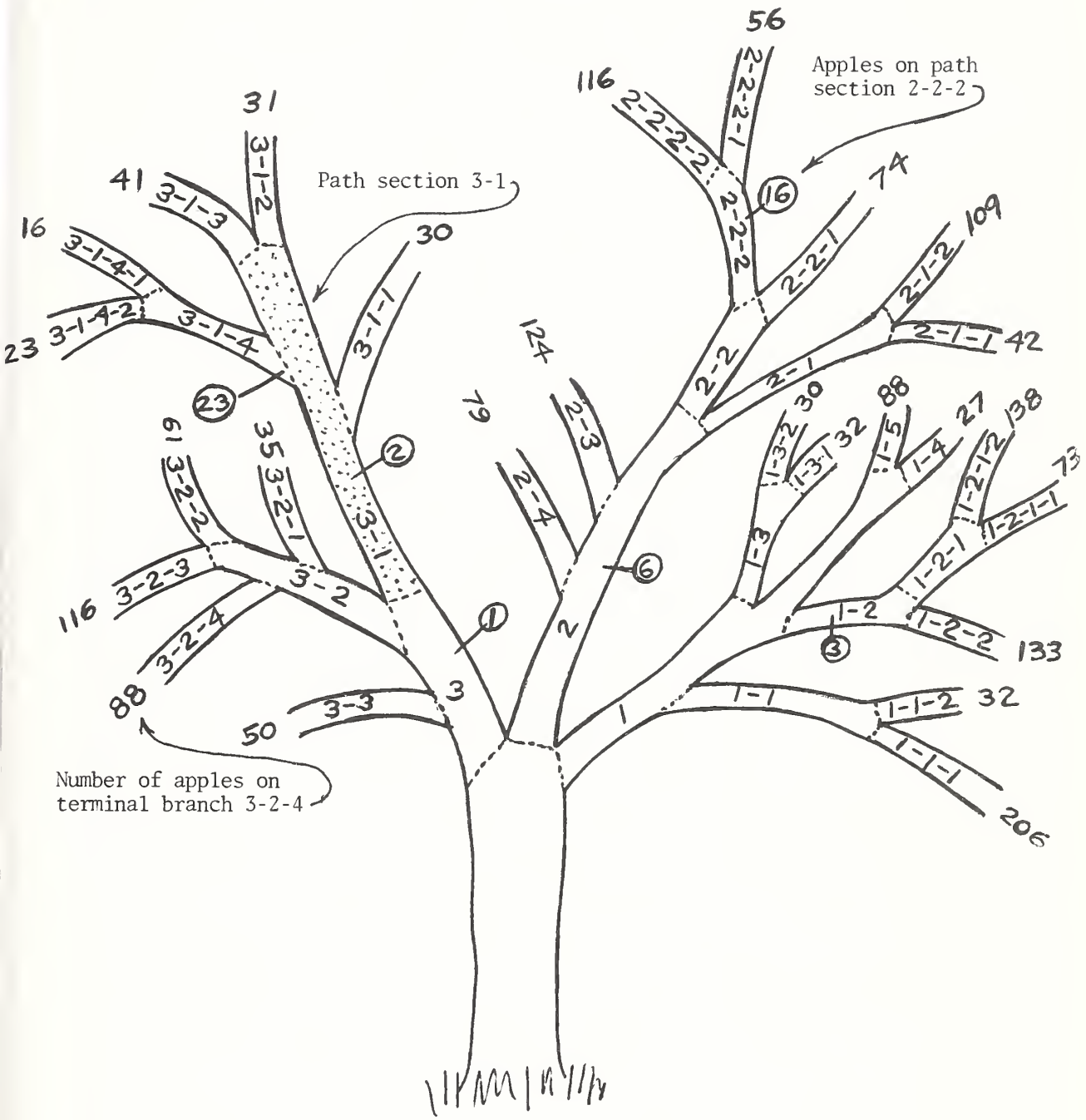


Figure 2.1--Map of apple tree no. 3 showing branch identification and number of apples

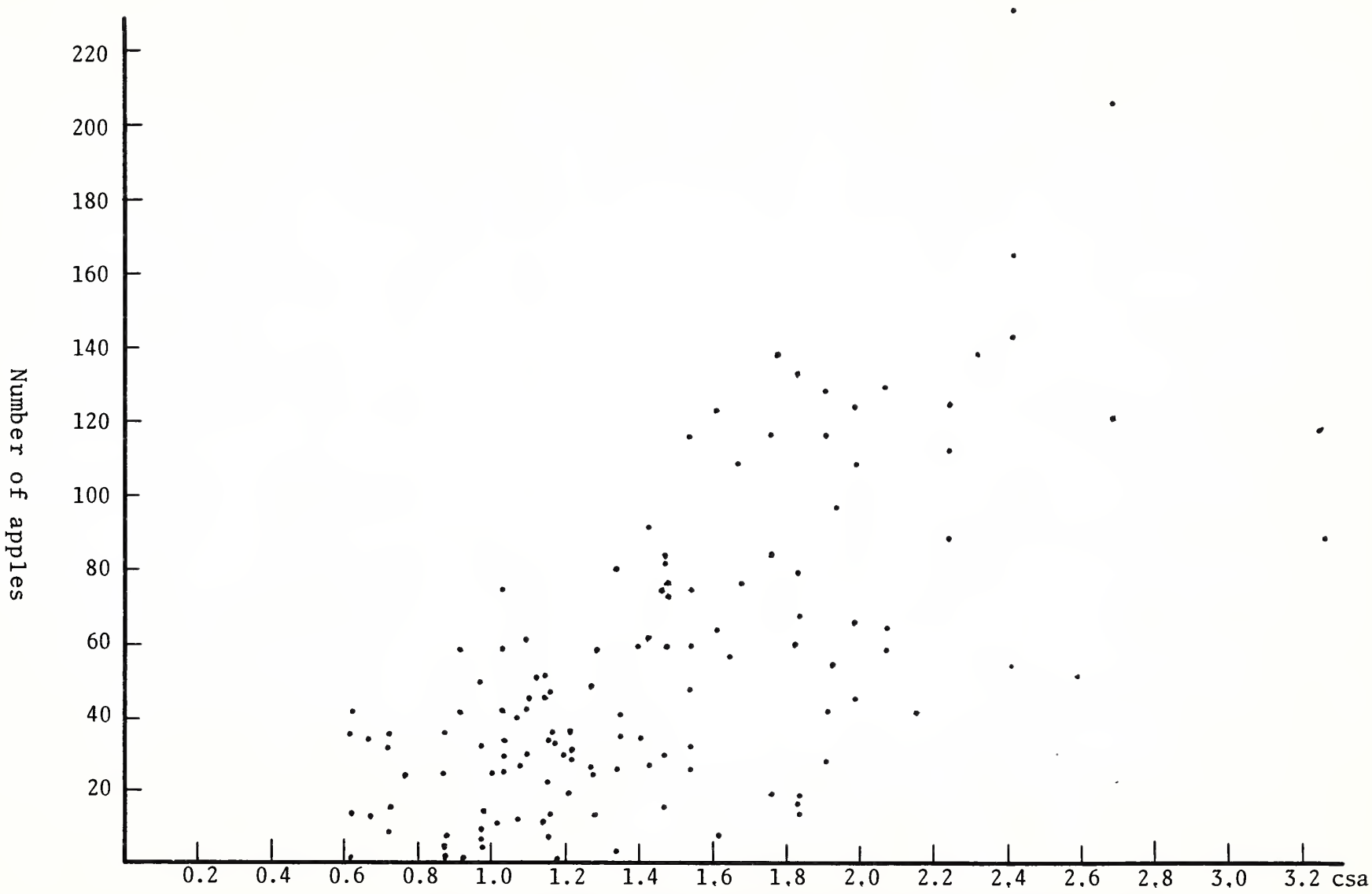


Figure 2.2--Number of apples by csa---terminal branches

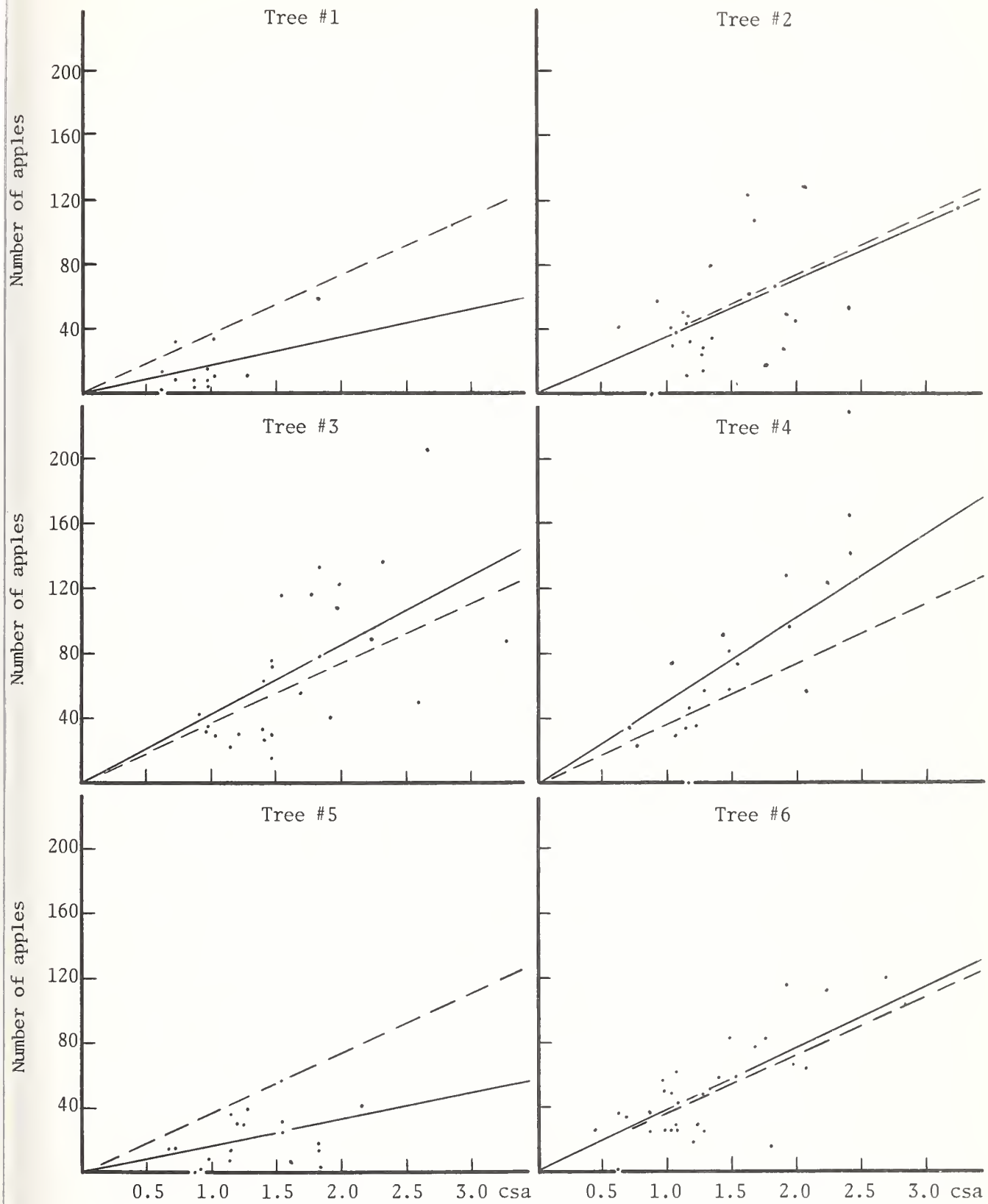


Figure 2.3--Dot charts and ratio lines by trees

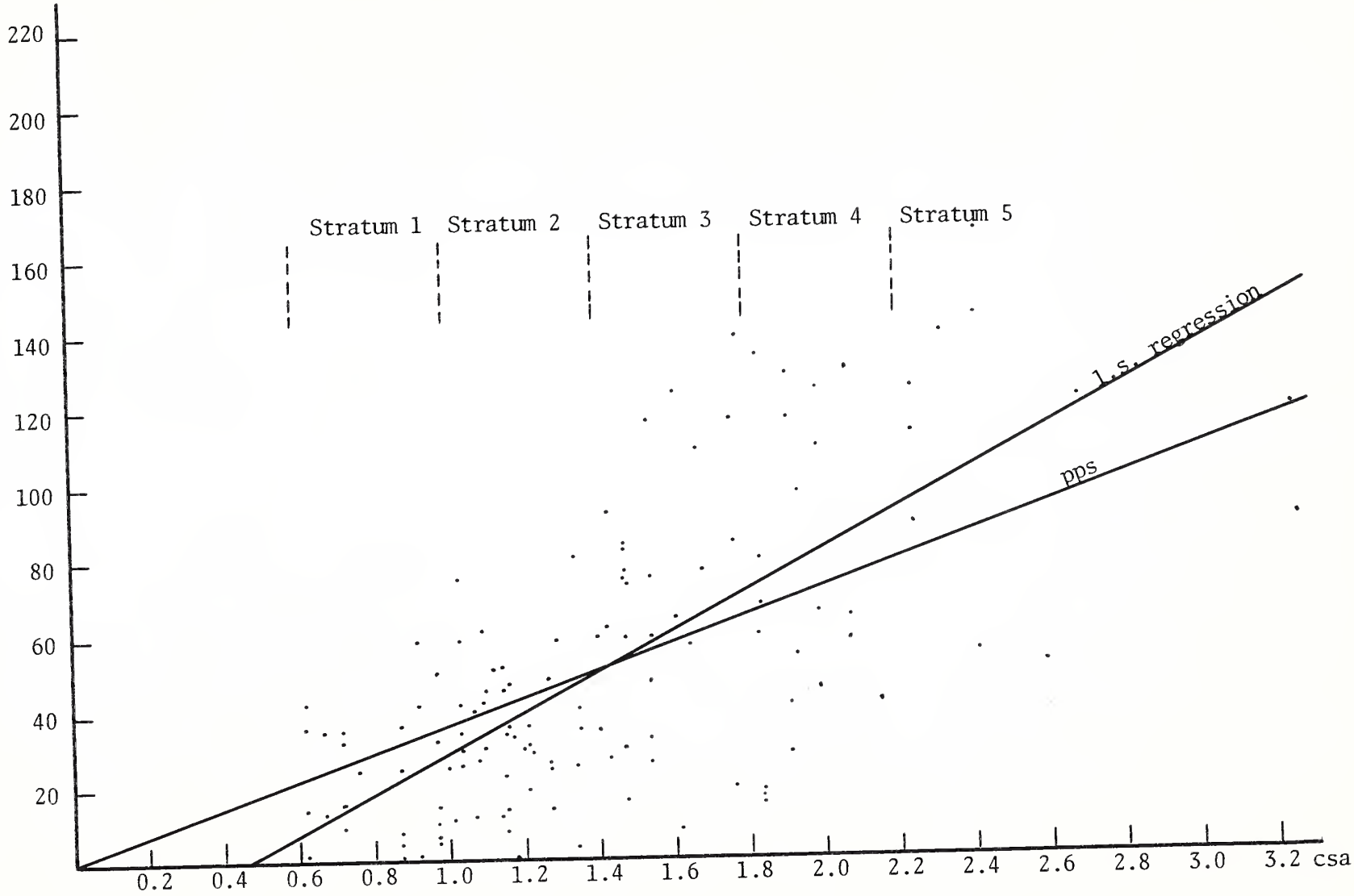


Figure 2.4--Size-of-branch strata, least squares regression line, and line for pps.

CHAPTER III

RANDOM-PATH SAMPLING OF FRUIT TREES

3.1 INTRODUCTION

The methods discussed in Chapter II required a map of each tree to be sampled. A map of a tree provides a good sampling frame, but drawing a map and measuring the csa's of all branches is too time consuming. In the research for practical ways of probability sampling, photographs of trees taken when the trees had no leaves have been studied for possible use as sampling frames, including the estimation of branch sizes for sampling with pps. Photography has also been considered, in the context of double sampling, as a means of counting and estimating the number of fruit on the tree.

In this chapter the random-path method proposed by Jessen[6/] for sampling fruit on a tree will be illustrated using one of the six apple trees in the analysis in Chapter II. Two random-path methods will be compared with two methods that were discussed in Chapter II. The comparisons will be made as though only one terminal branch from a tree is to be selected and used to estimate the total number of fruit on the tree.

6/ Jessen, Raymond J., Determining the Fruit Count on a Tree by Randomized Branch Sampling, Biometrics, March 1955.

3.2 FOUR METHODS OF SAMPLING A TREE

The four sampling methods and estimators, as described in this section, include only apples that are on terminal branches. A small proportion of the apples are not on terminal branches. Methods of including these apples will be discussed later.

(1) The first method is included primarily as a base for comparison. It is the same as Plan 1 discussed in Chapter II. After all terminal branches on a tree have been identified and numbered, one branch is selected at random with equal probability. Apples on the sample branch are then counted. The estimator is Ny_i where N is the number of branches on the tree and y_i is the number of apples on the sample branch. Since y_i refers to a sample value, one would expect i to be the index for branches in a sample. But, we are considering a sample of only one branch and it is convenient to let i be the index to branches in the population. Thus, if the 5th branch in the population, Y_1, Y_2, ..., Y_N, happens to be selected, $y_i = Y_5$. The first method will be referred to as DS-EP, which means direct selection from a list of all branches with equal probabilities.

(2) The second method is a random-path technique. Beginning from the bottom of the tree, the primary branches are all identified and one of the primary branches is selected at random with equal probabilities. The selected primary branch is then examined to identify all second-stage branches from it. Then, one second-stage branch is selected at random with equal probabilities. The process

is discontinued when a terminal branch has been selected. The estimator is $\frac{y_i}{p_i}$ where p_i is the probability of selecting the particular terminal branch that happens to be selected. As a short title for this method RP-EP will be used where RP represents random path and EP means equal probability of selection at each stage of branching.

(3) Like the first method, the third requires a complete identification of all terminal branches prior to selection. The csa (cross sectional area) of each terminal branch is measured and one branch is selected with pps, probability proportional to csa. The estimator is $X\frac{y_i}{x_i}$, where x_i is the csa of the selected branch and X is the sum of the csa's of all terminal branches on the tree. DS-PPS is the short title for this method, which is the same as plan 2 in Chapter II.

(4) The fourth method is a random-path method which differs from method two in the probability of selection. At each stage of branching the csa's of the branches at that stage are measured and one branch is selected with pps. The estimator, $\frac{y_i}{p_i}$, is like the estimator for the second method but the values of P_i are different. This method is titled RP-PPS.

3.3 BRANCH IDENTIFICATION AND DESCRIPTION OF DATA

Data for tree No. 3, which was represented in Figure 2.1, are presented in Table 3.1 in a way that shows the stage of branching. There were only three primary branches. Their csa's

11.60, 13.45, and 12.84, are presented in the column titled csa under 1st stage. The sum of these csa's is 37.89. Thus, if a primary branch is selected with pps, the first primary branch would have a selection probability equal to $\frac{11.60}{37.89}$. For further illustration of the recording system, notice that the second digit of the identification number shows four second-stage branches from the second primary branch. The csa's of these four branches and their sum are recorded under 2nd stage. This scheme of branch identification and recording is continued until a terminal branch is reached. When this occurs, the number of apples on a terminal branch is recorded to the right of its csa. Thus, Table 3.1 shows, for example, that branch 2-3 was a terminal branch with a csa equal to 1.99 square inches and that it had 124 apples on it. The numbers in parenthesis are numbers of "path" apples which will be discussed later.

Terminal branches were defined as branches having a csa between 3/4 and 2 square inches. Adherence to exact size is not possible. For example, the first terminal branch 1-1-1 has a csa equal to 2.68 and is large enough so an additional stage of branching was probably considered. If from 1-1-1 there were two branches of about equal size, those two branches would have been terminal branches. Probably 1-1-1 divided into several branches that were too small to be considered as terminal branches. As another case, suppose at the last stage of branching there is a branch with a csa of 1.5 square inches which is dying and clearly has no fruit. This branch could be shown on the map but marked for exclusion and not counted as a branch for sampling purposes.

Along a path from the base of a tree to a terminal branch there are some branches which are much too small to qualify as terminal branches. Fruit on small branches along a path to a terminal branch have been called path fruit and must be accounted for in some way. For example, on the path from the base of 1-2 to the bases of 1-2-1 and 1-2-2 there were three apples. These three apples are recorded in Figure 2.1 and in Table 3.1 next to the csa of branch 1-2. The counts of apples along the paths are shown in parenthesis in Table 3.1 to distinguish such apples from apples on the terminal branches. There are various ways of dealing with the path fruit; but, first let us examine the probability of selecting any given terminal branch with regard to each of the four methods.

3.4 PROBABILITY OF SELECTION AND ESTIMATION

With the DS-EP method each one of the 26 terminal branches has a selection probability equal to $\frac{1}{26}$. With DS-PPS the i^{th} terminal branch has a probability of selection equal to $\frac{X_i}{X}$ where X_i is its csa and $X=\Sigma X_i$. Calculating the probabilities for the random-path methods is more involved. For example, consider terminal branch 1-2-1-2 and RP-EP. In Table 3.1 notice that the numbers of branches at each stage on the path to terminal branch 1-2-1-2 are 3, 5, 2, and 2. Thus, the probability of selecting branch 1-2-1-2 is $(\frac{1}{3})\ (\frac{1}{5})\ (\frac{1}{2})\ (\frac{1}{2}) = \frac{1}{60}$, which is the product of the probabilities of selection at the four stages. With RP-PPS the product of corresponding probabilities

at the four stages is $(\frac{11.60}{37.90})$ $(\frac{5.61}{14.94})$ $(\frac{4.13}{5.96})$ $(\frac{2.32}{3.80})$ = .04862

An estimator for each of the four methods was presented above. However, all of the four estimators can be written in the same form,

$$\hat{Y}_i = \frac{y_i}{p_i} \tag{3.1}$$

where $\hat{Y}_i$ is the estimator and i = 1, 2, ..., N is the index to terminal branches on the tree. If the i^{th} branch happens to be selected, then $\frac{y_i}{p_i}$ is the estimate of the total number of apples on the tree (except that path apples are not included) where y_i is the number of apples on the i^{th} terminal branch and p_i is the probability of selecting it. The value of p_i depends on the method of sampling. In fact there are four sets of probabilities, one for each method. These four sets of values are presented in Table 3.2 in the columns headed P_1, P_2, P_3, and P_4. The columns headed $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$, and $\hat{Y}_4$ contain estimates of the total number of apples on the tree depending upon the terminal branch that happens to be selected. A discussion of these estimates follows.

Exercise 3.1 Refer to Table 3.1 and for each of the four sampling methods compute the selection probabilities for terminal branches 2-3 and 3-2-4. Compare your answers with the probabilities presented in Table 3.2.

To include the path apples there are at least two possibilities that might be considered with regard to the DS-EP and

DS-PPS methods: (1) Count all path apples and add the count to the estimate of the total number of apples on terminal branches that is obtained from a sample of terminal branches; (2) Define sampling units that are sections of path between the base of the tree and the terminal branches. Then select a sample of such sections to estimate the path fruit.

With the random-path methods, it is necessary to count the apples on each section of the path along the path to a terminal branch. Also, it is necessary to determine the probability that each section of the path had of being in the sample.

Since the RP-PPS method is of primary interest, it will be used to illustrate how the path fruit can be accounted for in the estimation process. Consider the three apples (see Table 3.1) on the path section 1-2 which is the section between the base of 1-2 and the third stage branches. To find the probability of these three being in the sample, consider repeated application of a random-path sampling method. These three apples will be in the sample whenever this path section is traversed. Therefore, under the RP-PPS method, the probability of this path section being in the sample is $(\frac{11.60}{37.89})\ (\frac{5.61}{14.94}) = .1150$ which gives $\frac{3}{.1150} = 26.1$ apples to be included in the estimate whenever this path section is traversed.

It is important to observe that a random path always ends with one and only one terminal branch. There are three terminal branches, 1-2-1-1, 1-2-1-2, and 1-2-2 that are connected

to the path section 1-2, and 26.1 would be included in the estimate that is made from the selected terminal branch that follows path section 1-2. There are no path apples other than the three that have been mentioned between the base of the tree and the three terminal branches that follow path section 1-2. Therefore, the estimated total number of apples in the event any one of the three terminal branches is selected would be:

Terminal Branch	Estimate
1-2-1-1	$\frac{3}{.1150} + \frac{73}{.03103} = 2379$
1-2-1-2	$\frac{3}{.1150} + \frac{138}{.04864} = 2863$
1-2-2	$\frac{3}{.1150} + \frac{133}{.03530} = 3794$

With the application of either random-path method, and assuming path fruit are recorded for each path section that is traversed, an estimator that includes path fruit can be written in generalized form. It appears to be complicated, but an illustration follows that should help clarify it. The estimator is:

$$\hat{Y}_i = \frac{y_{oi}}{p_{oi}} + \ldots + \frac{y_{ki}}{(p_{oi})\ldots(p_{ki})} + \ldots + \frac{y_{ti}}{(p_{oi})\ldots(p_{ti})} \qquad (3.2)$$

where $i = 1, 2, \ldots, N$ is an index of the terminal branches

$k = 0, 1, \ldots, t$ is an index of the path sections of the path to the i^{th} terminal branch (t is not constant, it depends on i),

y_{ki} = the number of path fruit on the k^{th} path section of the path from the base of the tree to the i^{th} terminal branch,

p_{ki} = the conditional probability of selecting the k^{th} path section of the path to the i^{th} terminal branch, given that the preceding path section has been selected.

When k = 0, the path section referred to is the part of the tree between ground level and the bases of the primary or first stage branches. Given that the tree is in the sample, p_{oi} = 1 which means that this path section is a part of the path to every terminal branch. Generally, y_{oi}, the number of fruit on this section of the tree, will be zero. When k = t, the k^{th} path section becomes a terminal branch, thus y_{ti} is the number of apples on the i^{th} terminal branch, and p_{ti} is the conditional probability of selecting the i^{th} terminal branch given that the path section that it is connected to has already been selected.

Suppose application of the RP-PPS method leads to terminal branch 1-2-1-1. In this case, k = 0, 1, 2, 3, 4 and the values of y_{ki} and p_{ki} are:

Path section	No. of apples on path section, y_{ki}	Conditional probability, p_{ki}
0	0	1
1	0	$\frac{11.60}{37.89} = .3061$
2	3	$\frac{5.61}{14.94} = .3755$
3	0	$\frac{4.13}{5.96} = .6930$
4	73	$\frac{1.48}{3.80} = .3895$

Substituting these values of y_{ki} and p_{ki} in the estimator, (3.2), gives:

$$\frac{(0)}{(1)} + \frac{(0)}{(1)(.3061)} + \frac{(3)}{(1)(.3061)(.3755)} + \frac{(0)}{(1)(.3061)(.3755)(.6930)}$$

$$+ \frac{(73)}{(1)(.3061)(.3755)(.6930)(.3895)} = 2379$$

Exercise 13.2 For the RP-EP and RP-PPS methods use the estimator (3.2) to obtain estimates corresponding to terminal branches 3-1-2, 3-1-4-1, and 3-3. Your results should agree with the estimates presented in Table 3.2.

There is an alternative view of the above method of including the path fruit which leads to the same answers. The idea is to prorate the path fruit to the terminal branches that follow the sections of the path where the path fruit are found. The prorating is done according to the probabilities of selection. Under the RP-EP method the three apples on the path section 1-2 would be prorated as follows:

Terminal branch	Prorated amount
1-2-1-1	$(\frac{1}{2})(\frac{1}{2})(3) = .75$
1-2-1-2	$(\frac{1}{2})(\frac{1}{2})(3) = .75$
1-2-2	$(\frac{1}{2})(3) = 1.50$
	Total = 3.00

Notice that $(\frac{1}{2})(\frac{1}{2})$, $(\frac{1}{2})(\frac{1}{2})$, and $(\frac{1}{2})$ are the conditional probabilities of selecting one of the three terminal branches, given that the path section 1-2 has already been selected. The conditional probabilities add to 1 which verifies that the method of prorating accounts for all of the path fruit.

If 1-2-1-1, for example, is the selected terminal branch, .75 is added to 73, the number of apples on 1-2-1-1. The estimate of the total number of apples on the tree is then obtained by dividing 73.75 by the probability of selecting 1-2-1-1 which is $(\frac{1}{3})(\frac{1}{5})(\frac{1}{2})(\frac{1}{2}) = \frac{1}{60}$. Thus, (60)(73.75) = 4425. The branch total, 73.75, (including the prorated amount) appears in Table 3.1 in the column titled EP, and the expanded total, 4425, appears in Table 3.2 in the column titled $\hat{Y}_2$.

Under the RP-PPS method, the system of prorating is the same but the probabilities are different. Thus,

Terminal branch	Prorated amount
1-2-1-1	$(\frac{4.13}{5.96})(\frac{1.48}{3.80})(3) = .81$
1-2-1-2	$(\frac{4.13}{5.96})(\frac{2.32}{3.80})(3) = 1.27$
1-2-2	$(\frac{1.83}{5.96})(3) = .92$
	Total = 3.00

The estimator $\hat{Y}_i$, Eq. (3.2), can be written in a form that corresponds to the idea of prorating path fruit to terminal branches. Let $p_i = (p_{oi})\ldots(p_{ti})$, which is the probability of selecting the i^{th} terminal branch. It follows that

$$\hat{Y}_i = \frac{y_i}{p_i} \tag{3.3}$$

where $y_i = [(p_{1i})\ldots(p_{ti})y_{oi}]+\ldots+[(p_{(k+i)i})\ldots(p_{ti})y_{ki}]+\ldots+[y_{ti}]$

Thus, y_i is the number of fruit "on" the i^{th} terminal branch including prorated amounts of path fruit. Assuming the RP-PPS method and terminal branch 1-2-1-1 as an example, the value of y_i is $(\frac{4.13}{5.96})(\frac{1.48}{3.80})(3) + 73 = 73.81$ and $\hat{Y}_i$ is $\frac{73.81}{.03103} = 2379$ which gives the same result that was obtained when Eq. (3.2) was used.

Table 3.2, columns headed $\hat{Y}_2$ and $\hat{Y}_4$, present estimates of the total number of apples on the tree for the RP-EP and RP-PPS methods and each of the possible random paths. These estimates were obtained by using the technique of prorating path fruit,

Eq. (3.3). That is, estimates of the total number of apples were obtained by dividing the values of Y_i (last two columns of Table 3.1) by the appropriate probabilities which are presented in Table 3.2, columns P_2 and P_4.

For comparison of the four methods we now need to decide how to include the path fruit for the DS-EP and DS-PPS methods. If the amount of path fruit is small, the best method might be to count all path fruit at the time the tree is mapped to determine terminal branches. In this case, assuming a sample of one terminal branch, the estimator, would be

$$\hat{Y}_i = Y' + \frac{y_i'}{p_i} \tag{3.4}$$

where Y' is the number of path fruit, y_i' is the number of fruit on the i^{th} terminal branch and p_i is the probability of selecting the i^{th} terminal branch. Alternatives are not considered in this illustration because, from a practical viewpoint, interest is in the random path methods. Thus, as a matter of expediency, the estimator (3.4) was used to obtain the estimates, $\hat{Y}_1$ and $\hat{Y}_3$, that are presented in Table 3.2 for the DS-EP and DS-PPS methods. Since only 51 apples out of 1901 were on path sections, the method of accounting for the apples on path sections probably has a very small impact on the sampling variance.

Exercise 3.3 For terminal branches 3-1-4-1 and 3-3, calculate estimates of the total number of apples on the tree for the DS-EP and DS-PPS methods using the estimator (3.4). Your answer should agree with the estimates that are presented in Table 3.1 for these two branches.

For each terminal branch and each of the four estimators (methods) there is a unique estimate of the total number of apples. All four estimators are unbiased. By definition, an estimator is unbiased if the expected (average) value of the estimates that might occur is equal to the population value. To find the expected value of an estimator, each estimate must be weighted by the probability of its occurence.

Exercise 3.4 For the RP-EP and RP-PPS methods, compute the expected value of the estimates presented in Table 3.2. The answer, except for rounding error, should be exactly 1901, which is the total number of apples on the tree.

3.5 VARIANCES OF THE ESTIMATORS

With reference to the theory of expected values, the variance of a random variable, Y, is the average of the squared deviations of Y from its expected (average) value. To be more specific, suppose Y is a random variable that can equal one of a set of values $Y_1, Y_2, \ldots, Y_N$ with probabilities $P_1, P_2, \ldots, P_N$ where $\Sigma P_i = 1$. By definition, the average value of Y is

$$\bar{Y} = E(Y) = \Sigma P_i Y_i$$

and the variance of Y, which is the average value of $(Y-\bar{Y})^2$, is

$$E(Y-\bar{Y})^2 = \sum_{i}^{N} P_i(Y_i-\bar{Y})^2$$

Exercise 3.5 Show that $\Sigma P_i(Y_i-\bar{Y})^2 = \Sigma P_i Y_i^2 - \bar{Y}^2$

Consider the estimator for the RP-PPS method. It is a random variable that can equal any one of the set of values in column $\hat{Y}_4$ of Table 3.2. The set of probabilities is presented in column P_4. By definition, the variance of the estimator (or estimates) is

$$(.05492)(3751-1901)^2+\ldots+(.06163)(814-1901)^2=800,194$$

or using the right hand side of the equation in exercise 3.5,

$$(.05492)(3751)^2+\ldots+(.06163)(814)^2-(1901)^2 = 800,194$$

The result, 800,194, is the sampling variance for the RP-PPS method when only one terminal branch is selected. If four terminal branches (or random paths) were selected with replacement, four estimates of the tree total would be computed, one for each branch, and the variance of the average of the four estimates would be $\frac{800,194}{4} = 200,048$.

The sampling variances (for a sample of one branch) are presented in Table 3.3 for each of the four methods and each of the six trees. The third tree is the one that was used above as an example. It is not expected that the four methods will always rank in the same order from one tree to another. However, the results illustrate some points that are of interest and importance.

3.6 DISCUSSION OF THE METHODS

The RP-EP method requires considerably less time than the RP-PPS method, but it has relatively high sampling variance because, at any given stage of branching, a large branch has the same probability of selection as a small one. That is, the RP-EP method is such that the probability of selecting a terminal branch has little or no relation to the number of fruit on the branch. The result, as shown by the sampling variances in Table 3.3, is a good illustration of a point that was made earlier. Compared to selecting sampling units with equal probability (as in the DS-EP method), the introduction of unequal probabilities of selection (as in the RP-EP method) will increase the sampling variance unless the selection probabilities are related to the values of the characteristic being measured in a way that will reduce sampling variance.

Figure 3.1 is a dot chart with the number of apples on a branch (column headed EP in Table 3.1) plotted against the values of P_2. The wide range in the selection probabilities and the lack of a relation explains the high sampling variance of the RP-EP method compared with the other methods. For comparison, Figure 3.2 is a dot chart for number of apples and the selection probabilities for the RP-PPS method. Compare Figure 3.2 with Figure 1.2 which showed a dot chart where sampling with pps would rank high.

After a branch has been identified and marked, the time required to obtain its csa, with a convenient instrument that

gives a reading directly in square inches (or square centimeters), is quite small. The use of csa as an auxiliary variable reduced sampling variance by a large amount. The reduction in variance in relation to cost is definitely advantageous. According to Table 3.3, the sampling variances for DS-PPS and RP-PPS are about the same and much less than the sampling variance for the DS-EP method. This indicates that RP-PPS is a good choice because it avoids the work of identifying all terminal branches before sampling. However, results in Table 3.3 should not be accepted as representative. The csa is not always an effective measure. Pruning and maintenance practices, age of trees, species or variety of trees, and other factors have some influence on the relation between csa and number of apples. The purposes of an intensive investigation limited to a few trees include testing different procedures for counting apples or measuring the size of branches, and acquiring ideas that seem to be worth exploring as possibilities for large scale application.

It is extremely important in the processes of sampling to understand the part played by randomization. Important biases sometimes occur even when strict attention is paid to details in making random selections. On the other hand, subjective evaluations or determinations in sampling are commonplace. With knowledge of how various factors effect sampling variance, the exercise of good judgement can be very effective in reducing sampling variance. But, there are points in the processes of

sampling where a determination should be strictly random. Some design constraints may be determined subjectively but selections of units for a sample should be in accord with rigorous, technical interpretation of randomness. It is generally preferable to have random selections made under competent supervision in an office, but that is not always feasible. Thus, one advantage of taking photographs of a sample of bare trees (assuming it is feasible) is that sample branches can be selected in the office. The selected branches are marked on photographs for enumerators. In this situation an enumerator's work is subject to full verification. Incidentally, the economics of sample surveys suggests that larger investments in sample design and selection can often be justified when the same sample is to be used for several surveys rather than one.

Exercise 3.6 Suppose the RP-PPS method is being applied and in the process you come to the following situation:

Assume that branch A, which has a csa equal to 3.2 square inches, has already been selected. It divides into two branches A-1 and A-2 with csa's equal to 1.4 and 1.6. With regard to size, the two branches, A-1 and A-2, qualify as terminal branches and ordinarily A-1 and A-2 would be accepted as terminal branches. But, before selecting one of the two, you happen to notice that A-2 has no apples on it and that A-1 appears to have approximately an average amount. Consider the following alternatives:

(1) *Accept A, which includes A-1 and A-2, as the terminal branch, and expand the count of apples by* $\frac{1}{P_A}$, *where* P_A *was the probability of selecting A.*

(2) *Accept A-1 and A-2 as terminal branches and select one with pps. Expand the count on A-1 or A-2 by* $(\frac{1}{P_A})(\frac{3.0}{1.4})$ *or* $(\frac{1}{P_A})(\frac{3.0}{1.6})$, *depending on whether A-1 or A-2 is selected.*

(3) *Discard A-2 since it has no apples on it and take A-1 as the terminal branch using* $(\frac{1}{P_A})(\frac{3.0}{1.4})$ *as the expansion factor.*

Discuss the alternatives with regard to bias and sampling variance.

Exercise 3.7 Refer to exercise 3.6 and as a variation of the situation assume that branch A-2 has been selected at random in accord with the instructions for the random path method. The enumerator prepares to count the apples on A-2 but finds there are no apples. He recognizes, since a sample of only one branch is to be selected for the sample from this tree, that the estimate of the number of apples on the tree will be zero (assuming no path fruit on the path to A-2). There is obviously a large number of apples on the tree, so he might have a strong opinion that something should be done that would give a better sample. How would you respond to each of the following possibilities:

(1) *Accept A-2 as a terminal branch, which means using zero as an estimate of the number of apples on the tree. Remember A-2 has already been selected.*

(2) Reject A-2 as a sample. Start at the beginning and select another terminal branch to replace A-2.

(3) Accept A which includes A-1 and A-2, as the terminal branch for the sample.

Discuss the three possibilities with regard to bias and sampling variance.

Exercise 3.8 In application of the RP-PPS method would it be advisable to be looking forward, as one approaches the terminal branch stage, for branches that are large enough to be terminal branches but clearly have a very small number of apples on them. With reference to the diagram in exercise 3.6 as an example, an enumerator looking forward, and considering what was ahead, could have stopped when A was selected and accepted A as a terminal branch. Otherwise, he would normally have followed the selection procedure one stage further. In application of the random path method, what is your opinion of the feasibility of looking ahead and taking eye estimates of numbers of apples into account in determining the terminal branch. Can it be used to reduce sampling error without risk of introducing bias? Think about the matter with regard to instructions that would be given to enumerators.

Exercise 3.9 It is not likely that there would be an interest in estimating the average number of terminal branches per tree. However, as an exercise, suppose the RP-PPS method is applied to the tree for which data are presented in Table 3.1. Assume that the following four terminal branches are selected as a sample:

1-1-2, 1-2-1-2, 2-4, and 3-2-1. From this sample, estimate the number of terminal branches on the tree. (The selection probabilities have already been computed, see Table 3.2). The parameter being estimated is 26. Ans. 33.4.

Exercise 3.10 Suppose a sample of 25 apple trees has been selected and that four enumerators have been trained in the application of the RP-PPS method. Assume that each enumerator, working independently and using the RP-PPS method, selects a sample of one terminal branch from each of the 25 trees. It is unlikely that enumerators will interpret terminal branches in exactly the same way. For example, one enumerator might have a tendency to follow the random path to terminal branches of the smallest permissible size, whereas another might stop as soon as he obtains a branch that is small enough to qualify as a terminal branch. Or, a branch along a path might be treated as a terminal branch by one enumerator and as path fruit by another. However, for each enumerator an estimate of the total number of apples on each tree is made using either (3.2) or (3.3) as the estimator. The 25 estimates are added together to obtain an estimate of the total number of apples on the 25 trees. This gives four estimates, one for each enumerator, of the total number of apples on the 25 trees.

(a) Assume that random selection is performed correctly at each stage of branching (after all branches at the stage have been completely identified and measured), and assume that apples have

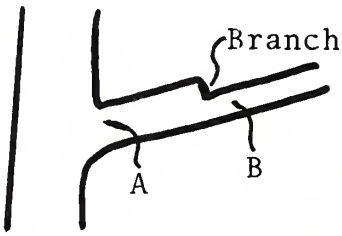
been correctly counted. Do the four estimates of the total number of apples all have the same expected value and the same variance?

(b) Suppose four estimates, one for each enumerator, of the total number of terminal branches on the 25 trees are made. Do these estimates have the same expected value? Why?

(c) Two enumerators measuring the csa's of any given set of branches are not likely to obtain exactly the same numerical values. Is this important? Discuss.

(d) The assumptions made in (a) are subject to question. Try listing some differences among enumerators that will, and will not, have an effect on the expected value of an estimate of the total number of apples on the 25 trees.

Exercise 3.11 Suppose, owing to pruning practices, that many cases like the following are found:



Assume the instructions were to always measure the csa at the base (point A) of a branch, Would you expect the csa measurements under the RP-PPS method to be ineffective, or even increase the sampling variance, compared with the DS-EP method? In cases like the above drawing, perhaps measuring the csa at position B would be more effective. What is your opinion? Incidentally, this is

a good example of why it is essential that a research and development staff should have actual experience with practical operations and decisions that must be made by enumerators. Do not expect high quality results when instructions are not well adapted. Agreement between concepts (the theoretical model) and operations as actually performed is of fundamental importance.

Table 3.1--Data by Branches for Apple Tree No. 3

Branch identification	1st stage		2nd stage		3rd stage		4th stage		No. of apples on or assigned to a terminal branch	
	csa	No.	csa	No.	csa	No.	csa	No.	EP	PPS
1-1-1	11.60		3.65		2.68	206			206	206
1-1-2					.97	32			32	32
					3.65					
1-2-1-1			5.61	(3)	4.13		1.48	73	73.8	73.8
1-2-1-2							2.32	138	138.7	139.3
							3.80			
1-2-2					1.83	133			134.5	133.9
					5.96					
1-3-1			2.01		.97	32			32	32
1-3-2					1.03	30			30	30
					2.00					
1-4			1.43	27					27	27
1-5			2.24	88					88	88
			14.94							
2-1-1	13.45	(6)	3.36		.92	42			42.8	42.5
2-1-2					1.99	109			109.7	110.1
					2.91					
2-2-1			5.09		1.47	74			74.7	74.7
2-2-2-1					3.47	(16)	1.64	56	64.4	65.2
2-2-2-2							1.54	116	124.4	124.6
					4.94		3.18			
2-3			1.99	124					125.5	125.0
2-4			1.83	79					80.5	79.9
			12.27							
3-1-1	12.84	(1)	6.30	(2)	1.47	30			30.6	30.4
3-1-2					1.21	31			31.6	31.3
3-1-3					1.91	41			41.6	41.5
3-1-4-1					4.13	(23)	1.47	16	27.7	29.6
3-1-4-2							1.15	23	34.8	33.6
					8.72		2.62			
3-2-1			5.35		1.40	35			35.1	35.1
3-2-2					1.42	61			61.1	61.1
3-2-3					1.76	116			116.1	116.1
3-2-4					3.26	88			88.1	88.1
					7.84					
3-3			2.59	50					50.3	50.2
	37.89		14.24						1901.0	1901.0

Total number of apples on terminal branches 1850
Total number of apples on path sections 51
Grand total 1901

Table 3.2 Probabilities of Selection and Estimates of the Total Number of Apples on Tree No. 3

Terminal branch no.	DS-EP		RP-EP		DS-PPS		RP-PPS	
	P_1	$\hat{Y}_1$	P_2	$\hat{Y}_2$	P_3	$\hat{Y}_3$	P_4	$\hat{Y}_4$
1-1-1	.03846	5407	.03333	6180	.06095	3431	.05492	3751
1-1-2	.03846	883	.03333	960	.02206	1502	.01988	1610
1-2-1-1	.03846	1949	.01667	4425	.03366	2220	.03103	2379
1-2-1-2	.03846	3639	.01667	8325	.05276	2667	.04864	2863
1-2-2	.03846	3509	.03333	4035	.04162	3247	.03530	3794
1-3-1	.03846	883	.03333	960	.02206	1502	.01998	1602
1-3-2	.03846	831	.03333	900	.02342	1332	.02121	1414
1-4	.03846	753	.06667	405	.03252	881	.02930	921
1-5	.03846	2339	.06667	1320	.05094	1776	.04590	1917
2-1-1	.03846	1143	.04167	1026	.02092	2059	.03073	1384
2-1-2	.03846	2885	.04167	2634	.04526	2459	.06647	1657
2-2-1	.03846	1975	.04167	1794	.03343	2265	.04382	1706
2-2-2-1	.03846	1507	.02083	3090	.03730	1552	.05334	1222
2-2-2-2	.03846	3067	.02083	5972	.03502	3363	.05009	2489
2-3	.03846	3275	,08333	1506	.04526	2791	.05757	2171
2-4	.03846	2105	.08333	966	.04162	1949	.05294	1509
3-1-1	.03846	831	.02778	1101	.03343	948	.02528	1203
3-1-2	.03846	857	.02778	1137	.02752	1147	.02081	1506
3-1-3	.03846	1117	.02778	1497	.04344	995	.03284	1264
3-1-4-1	.03846	467	.01389	2001	.03343	530	.03984	742
3-1-4-2	.03846	649	.01389	2505	.02615	931	.03117	1078
3-2-1	.03846	961	.02778	1263	.03184	1150	.02274	1542
3-2-2	.03846	1637	.02778	2199	.03229	1940	.02306	2648
3-2-3	.03846	3067	.02778	4179	.04003	2949	.02858	4062
3-2-4	.03846	2339	.02778	3171	.07414	1238	.05294	1665
3-3	.03846	1351	.11111	453	.05890	900	.06163	814
	.99996		1.00001		.99997		1.00001	

Table 3.3 Variances of Estimates of the Total Number of Apples on Each of Six Trees from a Sample of One Terminal Branch

Tree	No. of terminal branches	csa of trunk	No. of apples on tree	Variances			
				DS-EP	RP-EP	DS-PPS	RP-PPS
				(000)	(000)	(000)	(000)
1	13	7.0	214	40	28	24	22
2	27	20.0	1448	882	1383	674	478
3	26	23.0	1901	1419	2815	755	800
4	20	16.5	1658	1148	1444	380	350
5	19	13.5	403	82	263	65	79
6	30	19.5	1575	894	4339	416	513
Total	135	99.5	7199	4465	10272	2314	2242

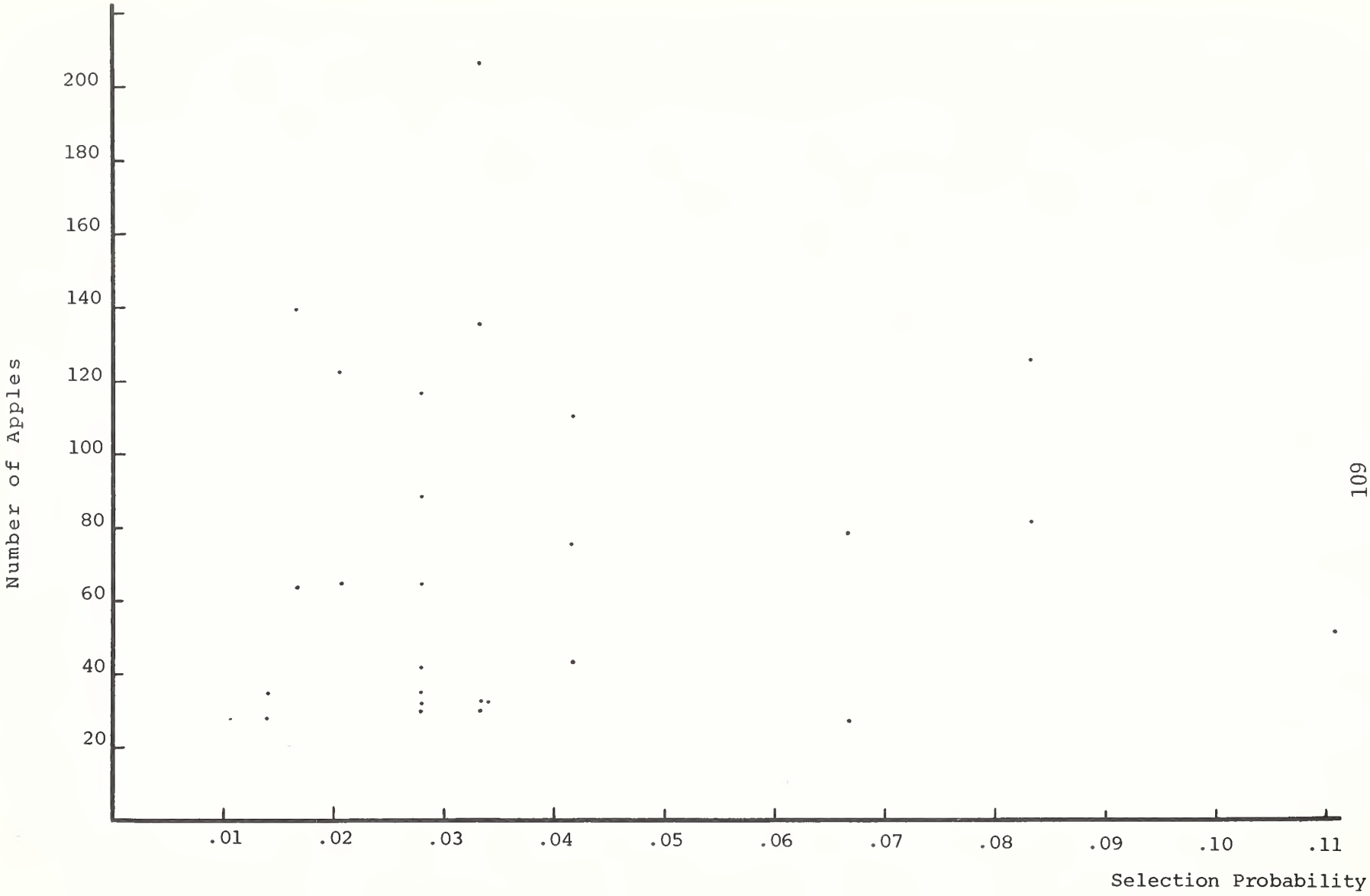


Figure 3.1 Dot Chart---Number of Apples vs Selection Probabilities for RP-EP

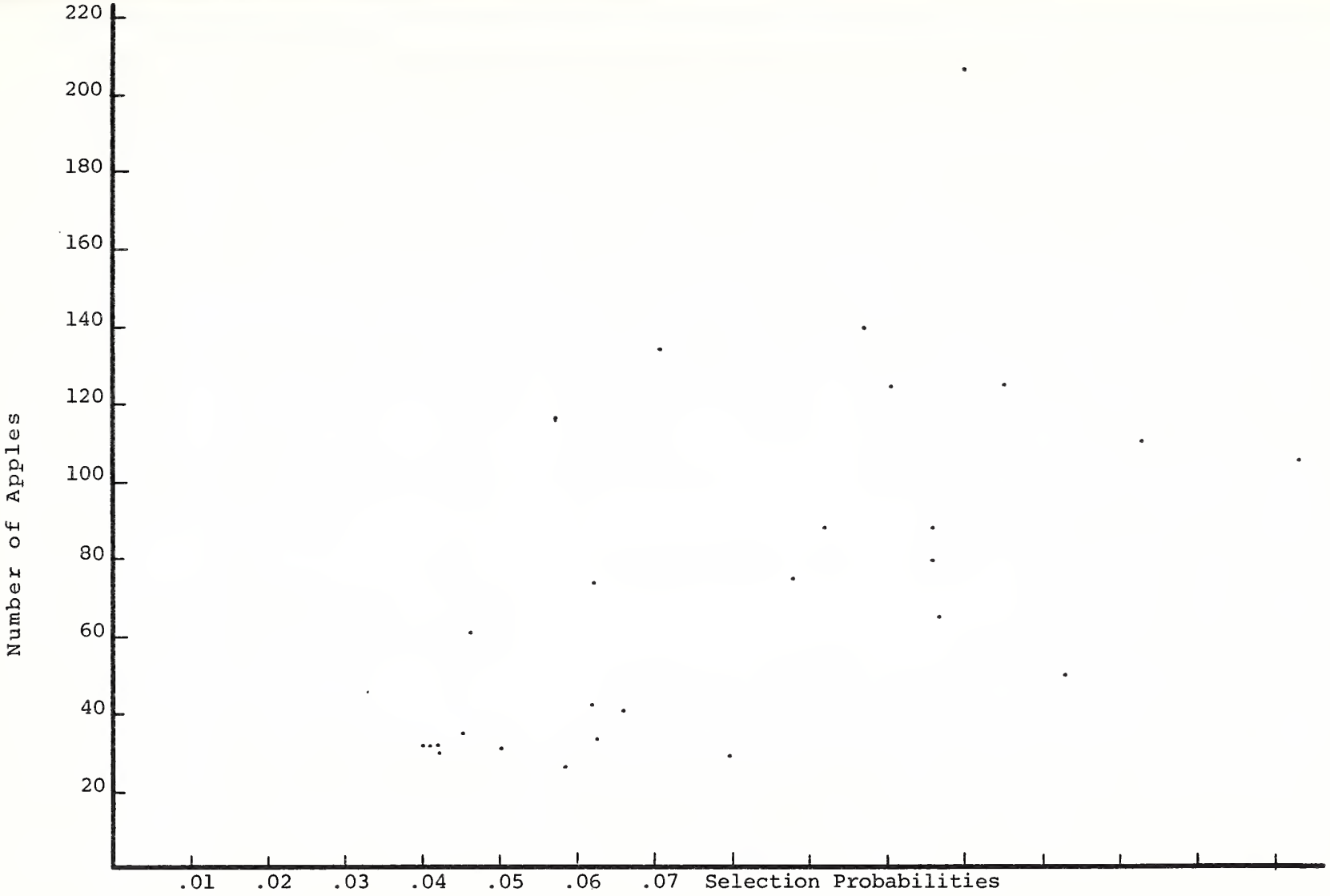


Figure 3.2 Dot Chart---Number of Apples vs Selection Probabilities for RP-PPS

TWO-STAGE SAMPLING

CHAPTER IV

4.1 INTRODUCTION

Most sampling plans for estimating or forecasting tree-crop production will involve three or four stages of sampling. Typically, there will be a sample of orchards (fields), a sample of trees in selected orchards, and a sample of branches from a sample of trees. Fruit on the sample branches would be counted and a small sample of fruit on the sample branches might be selected for measurements of size of fruit.

This chapter illustrates some alternative two-stage sampling plans using data for the six apple trees. Trees are the psu's (primary sampling units) and terminal branches or "paths" are the ssu's (secondary sampling units). The six trees will be treated as a population to be sampled and population variance formulas will be used to find the first and second-stage components of variance. Incidentally, the problem of making accurate counts of numbers of fruit on sample branches needs serious consideration. However, in the illustrations that follow, attention is limited to matters of sampling.

In the application of two-stage sampling, psu's are often selected with probabilities proportional to N_i, where N_i is the number of ssu's in the i^{th} psu. For some surveys, sampling with

probability proportional to N_i has important advantages. When the N_i are not known, approximations of N_i are often used.

With regard to sampling trees, the N_i (number of branches on trees) are not known and it is not feasible to determine the N_i for trees in an orchard. Some other effective measure of size must be found or the sample trees will need to be selected with equal probability. One possibility is to use a double sampling procedure. For example, a "large" sample of trees might be selected with equal probabilities. For each tree in the large sample a measurement of size, that takes relatively little time, might be made and used in the selection of a small sample of trees from the large sample. Possible measures of size are the csa of the trunk, the sum of the csa's of primary branches, and eye estimates of the amount of fruit. The feasibility of double sampling would depend upon the cost of obtaining the measurements of size and the relation between the measure of size and the amount of fruit on the trees. Stratification of trees within an orchard also needs to be considered. Sometimes strata within an orchard are readily recognized; for example, differences in age or variety. Perhaps a relation between size of trunk and number of apples will be found to be effective only within strata comprised of trees of the same variety and of a uniform condition.

Stratification, systematic sampling, or other techniques might be applied at any stage of sampling. However, for simplicity, the discussion will be limited to: (1) simple random sampling of psu's (selection with equal probability and without replacement)

and (2) sampling the psu's with pps (sampling with unequal probabilities of selection and replacement). Within each selected psu we will assume that a simple random sample of n_i ssu's is selected. The number of psu's in the sample is m and the number of ssu's in the sample is $n = \sum^{m} n_i$.

Refer to Table 4.1 for an exposition of the notation that will be used for representing data for a population. Examine the notation carefully. Sample data are represented in the same way except that lower case letters are used.

Since a general mathematical formulation of estimators and their variances is rather complex for two-stage sampling, we will procede from specific cases to more general description. The primary purpose of the next section is to present an elementary view of two-stage sampling.

4.2 PRIMARY SAMPLING UNITS EQUAL IN SIZE

The simplest case of two-stage sampling is one where all psu's have the same number of ssu's, where simple random is applied at both stages, and where the same number of ssu's is selected from each psu in the sample. In this case, and with reference to the notation in Table 4.1, the N_i all equal $\bar{N}$ and the n_i all equal $\bar{n}$. To summarize, the sampling plan under consideration is to select a simple random sample of m psu's from a population of M psu's and a simple random sample of $\bar{n}$ ssu's from each of the m psu's, which gives a total sample of $n = m\bar{n}$ ssu's.

For illustration a hypothetical population of 4 psu's with 5 ssu's in each is assumed. The 20 values of Y_{ij} are presented in the top part of Table 4.2. Deviations of Y_{ij} from $\bar{\bar{Y}}$ are also presented. In single-stage sampling, there is one component of variance, namely the variance of $(Y_{ij} - \bar{\bar{Y}})$ which in the illustration is 487.053.

In two-stage sampling, each deviation $(Y_{ij} - \bar{\bar{Y}})$ divides into two deviations as follows:

$$(Y_{ij} - \bar{\bar{Y}}) = (\bar{\bar{Y}}_i - \bar{\bar{Y}}) + (Y_{ij} - \bar{\bar{Y}}_i)$$

The values of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ are a set of deviations which reflect the variation among psu's and the values of $(Y_{ij} - \bar{\bar{Y}}_i)$ form the other set which reflects variation among ssu's within psu's. Turn to Table 4.2 and verify the deviations (components) $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ and $(Y_{ij} - \bar{\bar{Y}}_i)$. Notice that the between psu component, $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$, varies from one psu to another but is constant within a psu. There are only M different values of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ and selecting a sample of m psu's is equivalent to selecting a sample of m values of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$. Also, study the values of the within psu component, $(Y_{ij} - \bar{\bar{Y}}_i)$. It varies from one ssu to another within a psu, but its average value is zero for each psu. Therefore, these deviations reflect only variation within psu's. The second stage of sampling is equivalent to selecting $m\bar{n}$ of the deviations, $(Y_{ij} - \bar{\bar{Y}}_i)$

Now consider the variance of $\bar{\bar{y}}$, the mean of a two-stage sample. The difference between $\bar{\bar{y}}$ and $\bar{\bar{Y}}$ may be expressed as follows:

$$\bar{\bar{y}} - \bar{\bar{Y}} = \bar{\bar{d}}_1 + \bar{\bar{d}}_2$$

where $\bar{\bar{d}}_1$ is the average value of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ for the m psu's in the sample and d_2 is the average value of $(Y_{ij} - \bar{\bar{Y}}_i)$ for the $m\bar{n}$ ssu's in the sample.

Exercise 4.1 With reference to Table 4.2, suppose that psu's 1 and 3 are selected at the first stage and that ssu's 1 and 4 are selected within psu No. 1 and ssu's 3 and 5 are selected within psu No. 3. Find the values of $\bar{\bar{y}}$, $\bar{\bar{d}}_1$, *and* $\bar{\bar{d}}_2$. *Verify that* $\bar{\bar{y}} - \bar{\bar{Y}} = \bar{\bar{d}}_1 + \bar{\bar{d}}_2$. *Ans.* 34-43 = -9.4 + 0.4.

Since $\bar{\bar{d}}_1$ is the average of m random values of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ and $\bar{\bar{d}}_2$ is the average of $m\bar{n}$ random values of $(Y_{ij} - \bar{\bar{Y}}_i)$, it follows that $\bar{\bar{d}}_1$ and $\bar{\bar{d}}_2$ are random variables. It happens that $\bar{\bar{d}}_1$ and $\bar{\bar{d}}_2$ are independent. Therefore, the variance of $\bar{\bar{y}}$ is equal to the variance of $\bar{\bar{d}}_1$ plus the variance of $\bar{\bar{d}}_2$. From knowledge of the variance of the mean of a simple random sample, one might anticipate what the variances of $\bar{\bar{d}}_1$ and $\bar{\bar{d}}_2$ are and hence the formula for the variance of $\bar{\bar{y}}$ which is:

$$V(\bar{\bar{y}}) = \frac{M - m}{M} \frac{S_1^2}{m} + \frac{\bar{N} - \bar{n}}{\bar{N}} \frac{S_2^2}{m\bar{n}} \qquad (4.1)$$

where S_1^2 is the variance of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ and S_2^2 is the variance of the deviations $(Y_{ij} - \bar{\bar{Y}}_i)$. In this case, S_2^2 is a simple average of the within psu variances, S_{2i}^2, which is logical since the psu's are equal in size and are selected with equal probabilities. Moreover, the within psu sample size is constant.

For the illustration, values of S_1^2 and S_2^2 as functions of the deviations $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ and $(Y_{ij} - \bar{\bar{Y}}_i)$ are shown at the bottom of Table 4.2.

In practice, the two sets of deviations $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$ and $(Y_{ij} - \bar{\bar{Y}}_i)$, would not be computed. The variances, S_1^2 and S_2^2, could be calculated as follows:

$$S_1^2 = \frac{1}{\bar{N}^2} \left[\frac{\Sigma Y_i^2 - \frac{Y^2}{M}}{M - 1} \right] \tag{4.2}$$

$$S_2^2 = \frac{1}{M(\bar{N}-1)} \left[\Sigma\Sigma_{ij} Y_{ij}^2 - \frac{\Sigma Y_i^2}{\bar{N}} \right] \tag{4.3}$$

Exercise 4.2 Use Eq.'s 4.2 and 4.3 to find the values of S_1^2 *and* S_2^2 *in the numerical example. Explain why* $\bar{N}^2$ *appears as a divisor in Eq. 4.2.*

Exercise 4.3 For m=2 *and* $\bar{n}$=2 *find the variance of* $\bar{\bar{y}}$ *using Eq. 4.1. Ans.*118.9

Exercise 4.4 Show algebraically, that the right hand side of Eq. 4.3 is equal to $\frac{\Sigma_i S_{2i}^2}{M}$, *where* S_{2i}^2 *is the variance among ssu's within the* i^{th} *psu.*

One partial check on a variance formula is to determine whether it reduces to known formulas for special cases. Two special cases are of interest: (1) When m = M, two-stage sampling becomes stratified random sampling. That is, the psu's become strata. Observe, when m = M, that the first term on the right side of Eq. 4.1 vanishes and the second term becomes the variance for a stratified random sample of $\bar{n}$ units from each stratum (psu). (2) When $\bar{n} = \bar{N}$, two-stage sampling reduces to single-stage cluster sampling. In this case the last term in Eq. 4.1 vanishes, leaving the first

term which is the variance for a cluster sample where the clusters (sampling units) are the psu's.

Exercise 4.5 Suppose m=1 *and* $\bar{n}$=1. *In this case the selection of one psu at random and the selection of one ssu within it is equivalent to a single-stage sample of one ssu. Therefore, the variance of* $\bar{\bar{y}}$ *given by Eq. 4.1 when* m=1 *and* $\bar{n}$=1 *should be equal to the variance of* $\bar{\bar{y}}$ *for a single-stage random sample when* n=1. *Verify this using the data in Table 4.2. Remember the appropriate variance formula for the single-stage sample is Eq. 1.4.*

It is important to study the structure of the variance formula, Eq. 4.1, for the variance of $\bar{\bar{y}}$. When the number of psu's in the sample is fixed, increasing the size of the sample in each psu reduces only the second component of variance. As $\bar{n}$ increases, a point is reached where the among-psu variance is the major component and further increases in $\bar{n}$ contributes very little to reducing the variance of $\bar{\bar{y}}$. Notice that increasing m reduces both components when n_i is constant for all psu's.

4.3 PRIMARY SAMPLING UNITS UNEQUAL IN SIZE

Populations having psu's with equal numbers of ssu's are relatively infrequent. In this section, it is assumed that the numbers, N_i, of ssu's vary and that simple random sampling (without replacement) is applied at both stages.

As discussed in Chapter I, Sec. 1.1.2, "P" or "p" with appropriate subscripts refer to selection probabilities on the occasion of a particular random draw and "f" with an appropriate subscript refers to the probability that a particular unit has of being in the sample.

A general expression for the probability, f_{ij}, which any given ssu has of being included in a two-stage sample is:

$$f_{ij} = f_i f(j|i) \tag{4.4}$$

where f_i is the probability which the i^{th} psu has of being in the sample, and

$f(i|j)$ is the conditional probability which the j^{th} ssu in the i^{th} psu has of being in the sample, given that the i^{th} psu is in the sample of psu's.

With simple random sampling at both stages, $f_i = \frac{m}{M}$, and $f(j|i) = \frac{n_i}{N_i}$. Since f_i is constant for the case under consideration, let $f_i = f_1$ which is the sampling fraction at the first stage. Also, let $f(j|i) = f_{2i}$ which is the sampling fraction at the second stage within the i^{th} psu. Then Eq. 4.4 reduces to:

$$f_{ij} = f_1 f_{2i} \tag{4.5}$$

If the f_{2i} (the sampling fractions at the second stage) are constant, f_{ij} is constant and every ssu in the population has the same chance of being in the sample. Then, Eq. 4.5 becomes:

$$f = f_1 f_2$$

where f_2 is the constant second-stage sampling fraction. However, in the interest of generality, a requirement that f_{2i} be constant will not be specified at this point in the discussion.

An estimator of the population mean, $\bar{\bar{Y}}$, is

$$\hat{y} = \left(\frac{1}{N}\right)(M) \sum_{i}^{m} \frac{N_i \bar{\bar{y}}_i}{m} \tag{4.6}$$

where $\bar{\bar{y}}_i = \frac{\sum_j^{n_i} y_{ij}}{n_i}$ is the average of n_i ssu's in the sample from the i^{th} psu in the sample. Study the estimator 4.7 and observe that:

$N_i\bar{\bar{y}}_i$ is an estimate of Y_i, the total for the i^{th} psu;

$\sum_i^m \frac{N_i\bar{\bar{y}}_i}{m}$ is an average of the estimated totals for the m psu's in the sample; therefore,

$(M)\sum_i^m \frac{N_i\bar{\bar{y}}_i}{m}$ is an estimate of the population total and $(\frac{1}{\bar{N}})$ in Eq. 5.6, changes the estimated total to an estimate of $\bar{\bar{Y}}$.

The variance of $\hat{y}$ is given by:

$$V(\hat{y}) = \frac{1}{m}\left[(1-f_1)S_1^2 + \frac{1}{N^2}\sum_i^M M(1-f_{2i})\frac{N_i^2S_{2i}^2}{n_i}\right] \qquad (4.7)$$

where $S_1^2 = \frac{1}{\bar{N}^2}\frac{\sum_i^M (Y_i - \bar{Y})^2}{M-1}$ is the variance among psu totals divided by $\bar{N}^2$ so S_1^2 will be expressed on the basis of one ssu, and

$S_{2i}^2 = \frac{\sum_j^{N_i} (Y_{ij} - \bar{\bar{Y}}_i)^2}{N_i - 1}$ is the variance among ssu's within the i^{th} psu.

The first part of 4.7, $\frac{1}{m}(1-f_1)S_1^2$, is the variance of $\hat{y}$ assuming all of the m psu's are enumerated completely. That is, the theory for single-stage sampling applies to the first stage.

The quantity:

$$(1 - f_{2i}) \frac{N_i^2 S_{2i}^2}{n_i}$$

in Eq. 4.7 is recognizable as the variance of $N_i \bar{\bar{y}}_i$ where $\bar{\bar{y}}_i$ is the mean of a simple random sample of n_i ssu's in the i^{th} psu.

Eq. 4.7 was written in the above form for comparison with other variance formulas given later for two-stage sampling. The second term within [ ] could be written as follows:

$$\left(\frac{1}{\bar{N}^2}\right)\left(\frac{1}{M}\right) \sum_i^M (1 - f_{2i}) \frac{N_i^2 S_{2i}^2}{n_i} \qquad (4.8)$$

because $\frac{M}{N^2} = \frac{1}{\bar{N}^2} \frac{1}{M}$. Expression 4.8 shows that the variances of $N_i \bar{\bar{y}}_i$ are summed over all psu's in the population and the sum is divided by M giving an average of such variances. The variances of $N_i \bar{\bar{y}}_i$ receive equal weight in the average because the psu's are selected with equal probabilities. Since the average variance of $N_i \bar{\bar{y}}_i$ pertains to psu totals, the divisor $\bar{N}^2$ appears in 4.8 to convert the variance to a basis of one ssu. Such analysis of a formula is helpful in determining whether one has the right formula for a particular purpose.

Exercise 4.6 If the variance formulas (4.1) and (4.7) are correct, formula (4.7) should reduce to (4.1) when $N_i = \bar{N}$ *and* $n_i = \bar{n}$. *Show that this is true.*

When the second-stage sampling fractions $\frac{n_i}{N_i}$, are constant and equal to f_2, the estimator, (4.6), reduces to:

$$\hat{y} = \frac{\Sigma\Sigma y_{ij}}{f_2 m\bar{N}} \tag{4.9}$$

and its variance, (4.7), reduces to:

$$V(\hat{y}) = (1 - f_1)\frac{S_1^2}{m} + (1 - f_2)\frac{S_2^2}{m\bar{n}} \tag{4.10}$$

where S_1^2 is the same as in 4.7,

$$S_2^2 = \sum_i^M \frac{N_i}{N} S_{2i}^2,$$

and

$$\bar{n} = \frac{\sum_i^M n_i}{M} = \frac{\Sigma f_2 N_i}{M} = f_2\bar{N}$$

Exercise 4.7. Show that Eq.'s (4.9) and (4.10) follow from (4.6) and (4.7) when $f_2 = \frac{n_i}{N_i}$

Exercise 4.8 Show that $f_2 m\bar{N}$, *in Eq. 4.9, is equal to the expected sample size. That is, show that* $E(n) = f_2 m\bar{N}$ *where* $n = \sum_i^m n_i$. *In practice one would probably use* n, *the actual sample size, in the estimator instead of the expected size,* $f_2 m\bar{N}$. *Moreover,* $\bar{N}$ *is not known in most practical applications.*

4.3.1 NUMERICAL EXAMPLE

As a numerical example, the apple tree data presented in Table 2.1 will be treated as a population to be sampled. The psu's are trees and ssu's are terminal branches. The number of trees in an orchard is usually large and in practice the number of sample trees selected from an orchard would be relatively small, that is $(1 - f_1)$ would be nearly equal to 1. Accordingly, for this illustration, $(1 - f_1)$ is assumed to be 1 even though $M = 6$ and $(1 - f_1) = \frac{M-m}{M}$ is considerably less than 1.

Suppose we are interested in knowing what the sampling variance is for the following three allocations of a sample of four terminal branches assuming simple random sampling at both stages:

Allocation	No. of Trees m	No. of Branches Selected from Each Tree $n_i = \bar{n}$
1	1	4
2	2	2
3	4	1

To find the variances for the three allocations we need part of the results in Table 2.6. The relevant results, N_i, Y_i, and S_{2i}^2, from Table 2.6 are included in Table 4.3 along with some other information that will be used later.

In each allocation, n_i is constant (the same for all trees) which means that $\frac{n_i}{N_i}$ is not constant and the branches do not have equal probability of being in the sample. Thus, the estimator, Eq. 4.6, and its variance, Eq. 4.7, are applicable. The variances for the three allocations are presented in Table 4.4.

Exercise 4.9 Refer to the data presented in Table 4.3, columns N_i, Y_i, *and* S_{2i}^2 *and perform the calculations that are needed to obtain the results presented in Table 4.4 for* $m = 2$ *and* $n_i = \bar{n} = 2$. *Assume that* f_1 *is negligable.*

Exercise 4.10 Complete the following table:

				Variance Components	
n	*m*	$\bar{n}$	$V(\hat{y})$	*Among psu's*	*Within psu's*
1	1	1	1167.0		
2	1	2			
4	1	4			
2	2	1			
4	2	2			
8	2	4			
4	4	1			
8	4	2			
16	4	4	306.5		

If you understand the variance formula 4.7 and the results in Table 4.3, this table can be completed very easily. First, fill in the "Among psu's" column by copying the appropriate numbers from Table 4.4. Consider how to fill in the "Within psu's" column by making simple changes in the within psu components in Table 4.4. Study the results. For a constant value of $\bar{n}$ *and an increase in m from 1 to 4 there is a 75 percent reduction in the variance of* $\hat{y}$; *but, for a constant m, increasing* $\bar{n}$ *from 1 to 4 reduces the variance of* $\hat{y}$ *by less than 50 percent.*

Exercise 4.11 One of the numbers in Table 4.4 is the sampling variance for m = 2 and $n_i = N_i$. What is the number?

Exercise 4.12 Find the probability that any given terminal branch on tree No. 1 has of being in the sample when m = 2 and n_i = 2 for all trees. What is the probability for tree No. 3? Is the unequal probability something to be concerned about? In what ways?

It is of interest to compare the variance for a simple random (single-stage) sample of 4 branches with the variances of $\hat{y}$ in Table 4.4. The variance among the 135 branches is 1,762 (see Table 2.6). Hence, the variance of the mean of a sample of 4 branches is $\frac{1762}{4}$ = 440, disregarding the fpc. The answer, 440, is less than the variances of $\hat{y}$ in Table 4.4. This is expected with the possible exception of the allocation m = 4 and n_i = 1, which has a variance equal to 583.5. However, when one recognizes in the specified two-stage plans that all branches do not have the same probabilities of selection, it is resonable to expect that the answer for simple random sampling would be less than 583.5.

Suppose we wish to give every branch an equal chance of being in the sample. Considering samples of 4 branches the overall sampling fraction would be $\frac{4}{135}$. If we specify that m = 2, then $f_1 = \frac{1}{3}$ and all $\frac{n_i}{N_i}$ (or f_2) should equal $\frac{4}{45}$. Since the N_i are small and the n_i must be integers, it is not possible to have all $\frac{n_i}{N_i}$ exactly equal to $\frac{4}{45}$. This presents a type of practical problem that often occurs when working with small integers. Ways

of dealing with this problem will not be discussed at this point. Instead, we will procede as though the fraction $\frac{n_i}{N_i}$ is sufficiently close to $\frac{4}{45}$ to warrant use of the unweighted average of the sample data as the estimator and the variance formula 4.10. Assuming $(1 - f_1) = 1$, for reasons explained above, and substituting the numerical values of S_1^2 and S_2^2 in 4.10, we have:

$$V(\hat{y}) = \frac{1}{m}\left[917.1 + (1 - f_2)\frac{1367}{\bar{n}}\right] \qquad (4.11)$$

When $m = 2$ and $f_2 = \frac{4}{45}$, the value of $\bar{n}$ is 2 and the variance of $\hat{y}$ is 769.9. This answer compares with 797.6 in Table 4.4.

Exercise 4.13 Verify the numbers, 917.1 and 1367, in Eq. 4.11.

It is often desirable to specify that all ssu's in the population have an equal chance of being in the sample. As discussed above, one way of fulfilling this requirement is to select psu's with equal probability and apply a constant sampling fraction at the second stage of sampling. But, when the sizes of the pus's vary widely, this method often has two important disadvantages: (1) Variance associated with variation in the sizes of the psu's is included in the variance of an estimate unless such variation is reduced by design. Notice that S_1^2 in 4.7 is the variance among psu totals rather than the variance among psu means. Incidentally, an auxiliary variable(s) might be useful in reducing the sampling variance associated with the first stage of sampling. (2) When the second-stage sampling

fraction, $\frac{n_i}{N_i}$, is constant, n_i is proportional to N_i and the workload varies from one psu to another. For many surveys, it is important for reasons of economy that n_i, rather than $\frac{n_i}{N_i}$ be constant. Selecting psu's with pps is often very helpful in overcoming these disadvantages.

Exercise 4.14 Under the plan of applying a sampling fraction of $\frac{4}{45}$ *to each tree that is selected, suppose that trees numbered 1 and 3 are selected. Find the values of* n_i *for these two trees where* n_i *is* $\frac{4}{45} N_i$ *rounded to the nearest integer. Also, find* $n = \Sigma n_i$. *Do the same assuming trees numbered 2 and 4 are selected. This illustrates that the size of the sample,* $n = \Sigma n_i$, *is a random variable. Also, in this case,* $\frac{n_i}{N_i}$ *cannot be exactly constant. One should consider whether there is an appreciable bias in the estimator (4.9). Use (4.6) instead of (4.9) unless there is assurance that any bias in (4.9), owing to unequal probabilities of the ssu's being in the sample, is negligible.*

4.4 SELECTION OF PSU'S WITH PPS

Consider a sample of m psu's selected with replacement and with selection probabilities $P_1, P_2 \ldots, P_n$ (See section 1.1.2 in Chapter I). Let n_i be the size of a simple random sample of ssu's that is to be selected from the i^{th} psu in the event that it is selected. If, by chance, the i^{th} psu is selected a second time another sample of n_i ssu's is selected. For a sample of n psu's the estimator is:

$$\hat{\bar{y}} = \left(\frac{1}{N}\right)\left(\frac{1}{m}\right) \sum_{i}^{m} \frac{N_i \bar{y}_i}{p_i} \qquad (4.12)$$

Remember to interpret "i" as an index of the psu's selected by the m random draws. Notice that $N_i\bar{\bar{y}}_i$ is an estimate of a psu total and that $\frac{N_i\bar{\bar{y}}_i}{p_i}$ is an estimate of the population total, Y, based on a sample of one psu and a simple random sample of n_i ssu's within it. Thus, there are m estimates of Y and $(\frac{1}{m})\sum_i^m \frac{N_i\bar{\bar{y}}_i}{p_i}$ is an average of these estimates. The factor $\frac{1}{N}$ makes $\hat{y}$ an estimator of $\bar{\bar{Y}}$. The variance of $\hat{y}$, in Eq. 4.12, is:

$$V(\hat{y}) = \frac{1}{m}\left[\sigma_1^2 + \frac{1}{N^2}\sum_i^M (\frac{1}{P_i})(1-f_{2i})\frac{N_i^2 S_{2i}^2}{n_i}\right] \qquad (4.13)$$

where $$\sigma_1^2 = \frac{1}{N^2}\sum_i^M P_i(\frac{Y_i}{P_i} - Y)^2$$

and $$S_{2i}^2 = \frac{\sum_j^{N_i} (Y_{ij} - \bar{\bar{Y}}_i)^2}{N_i - 1}$$

Exercise 4.15 Compare $\frac{\sigma_1^2}{m}$ *in 4.13 with the variance of* $\hat{y}_4$ *in Table 1.1, using the alternative expression for* σ_4^2 *in the variance of* $\hat{y}_4$. *Change the notation used in Chapter 1 to conform to the notation used for psu's. This gives:*

$$V(\hat{y}_4) = (\frac{1}{m})(\frac{1}{M^2})\sum_i^m P_i(\frac{Y_i}{P_i} - Y)^2$$

Why is this expression for $V(\hat{y}_4)$ *different from the between psu part of the variance in 4.13? In terms of the notation for two-stage sampling* $\hat{y}_4$ *is an estimate of* Y *rather than* $\bar{\bar{Y}}$. *Change* $\hat{y}_4$ *so it will be an estimator of* $\bar{\bar{Y}}$ *and make the corresponding change in* $V(\hat{y}_4)$. *Your answer should agree exactly with* $\frac{\sigma_1^2}{m}$ *in (4.13).*

Notice the correspondence between S_1^2 in Eq. 4.7 and the variance of $\hat{y}_1$, plan 1, in Chapter I; also, notice the correspondence between σ_i^2 in Eq. 4.13 and the variance of $\hat{y}_4$, plan 4, in Chapter I. The discussion in Chapter I of the efficiency of plan 4 compared to plan 1 is relevant to the first stage of sampling. If N_i is a good measure of size, σ_1^2 will be considerably less than S_1^2.

Compare the components of variance in Eq. 4.7 and Eq. 4.13 which pertain to the second stage of sampling. The only difference is a reflection of the difference in the probabilities of selection at the first stage. When the probabilities are equal, $P_i = \frac{1}{M}$ and substituting $\frac{1}{M}$ for P_i in 4.13 gives 4.7.

In Eq. 4.4, f_{ij} was expressed as the probability that any given ssu has of being in a sample assuming the sample at both stages was simple random sampling without replacement. This equation now needs modification to be in accord with sampling at the first stage with unequal probability and with replacement. An appropriate probability equation is:

$$f'_{ij} = P_i f_{2i} = P_i \frac{n_i}{N_i} \tag{4.14}$$

where P_i is the selection probability, at any given random draw, for the i^{th} psu in the population,

f_{2i} as defined before, is the sampling fraction within the i^{th} psu of the population, and

f_{ij} is the probability which the j^{th} ssu in the i^{th} psu of the population has of being in a sample obtained by selecting <u>one</u> psu with pps and selecting a simple random sample of n_i ssu's within the selected psu.

It is in the context of the probability Eq. 4.14 that the estimator, 4.12 and its variance, 4.13, are applicable, assuming m independent random selections of psu's.

The estimator, Eq. 4.12, and its variance, Eq. 4.13, are for any given set of selection probabilities at the first stage and any given set of sample sizes, n_i, at the second stage. An important special case exists when f'_{ij}, in Eq. 4.14 is held constant and when the psu's are selected with probabilities proportional to N_i, that is, when $P_i = \frac{N_i}{N}$. By letting f' be the constant value of f'_{ij}, we obtain the following results from Eq. 4.14:

$$n_i = f'N = \bar{n}$$

and

$$f_{2i} = \frac{\bar{n}}{N_i}$$

That is, the sample size within a psu is constant, and, since f'_{ij} is also constant, the sample is self-weighted.

The estimator and its variance become:

$$\hat{y} = \frac{\Sigma\Sigma y_{ij}}{n} \tag{4.15}$$

and $$V(\hat{y}) = \frac{1}{m}\left[\sigma_1^2 + \frac{1}{\bar{n}}\frac{1}{N}\sum_i^M N_i(1 - f_{2i})S_{2i}^2\right] \tag{4.16}$$

where $$\sigma_1^2 = \frac{1}{N}\sum_i^M N_i(\bar{\bar{Y}}_i - \bar{\bar{Y}})^2$$

For computational purposes one might use:

$$\sigma_1^2 = \frac{1}{N}\sum_i^M \frac{Y_i^2}{N_i} - \left(\frac{Y}{N}\right)^2$$

and $$\frac{1}{N}\sum_i^M N_i(1 - f_{2i})S_{2i}^2 = \frac{1}{N}\Sigma N_i S_{2i}^2 - \frac{\bar{n}}{N}\Sigma S_{2i}^2$$

Exercise 4.16 Show that Eqs. 4.12 and 4.13 reduce to 4.15 and 4.16 when $P_i = \frac{N_i}{N}$ *and* $n_i = \bar{n}$.

When the N_i are not known, estimates of N_i or a suitable measure of size might be used in place of N_i. In this case, assuming $f'_{ij} = f'$, the sampler would choose a value of f' such that f'N is the desired average size of sample from a psu. Since the selection probabilities for psu's are known, the second-stage sampling fraction $f_{2i} = \frac{f'}{P_i}$ would be calculated for each selected psu. Application of these second-stage sampling fractions gives a self-weighted sample. The n_i will be nearly equal if the measure of size is close to being proportional to N_i. The estimator, Eq. 4.12, and its variance, Eq. 4.13, are applicable. They could be modified by making use of the fact that $p_i f_{2i} = p_i \frac{n_i}{N_i} = f'$.

4.4.1 NUMERICAL EXAMPLE

Exercise 4.17 With reference to the apple tree example, we found for simple random sampling at both stages that the sampling variance was 797.6 *when* m=2 *and* $n_i=\bar{n}=2$ *(See Table 4.4). For comparative purposes, find the sampling variance for* m=2 *and* $\bar{n}=2$ *when the trees are selected with probabilities proportional to* N_i. *The data needed are found in Table 4.3, columns headed* N_i, Y_i, *and* S_{2i}^2. *Find the values of* σ_1^2, $\frac{1}{N}\Sigma N_i S_{2i}^2$, *and* $\frac{1}{N}\Sigma S_{2i}^2$, *then compute the variance of* $\hat{y}$ *for* m=2 *and* $\bar{n}=2$. *Ans.* 532.6.

Substituting results from exercise 4.16 in Eq. 4.16 gives:

$$V(\hat{y}) = \frac{439.7}{m} + \frac{1367 - \bar{n}\ (57.99}{m\bar{n}} \qquad (4.17)$$

For m=2 the between psu variance, $\frac{439.74}{2}$ = 219.9, compares with 458.6 (see Table 4.4) when two psu's are selected with equal probability. As indicated by this result, selecting psu's with pps is often very important in reducing the between psu component of variance. For m=2 and $\bar{n}=2$ the within psu component in Eq. 4.17 is equal to 312.8 which compares with two other results that were obtained when the trees are selected with equal probabilities: 339.0 when n_i = 2, and 311.4 when $\frac{n_i}{N_i}$ is constant and $\bar{n}=2$. The first result, 339.0, was recorded in Table 4.4 and the second, 311.4, is readily obtained by Eq. 4.11.

Suppose that one tree is selected with probability $\frac{N_i}{N}$ and that one branch is selected from it with equal probability. In this case, m=1, and $\bar{n}$=1, and the variance of $\hat{y}$ according to variance formula 4.17 is 1748.8. The probability of selecting any given branch in the population is $(\frac{N_i}{N})(\frac{1}{N_i})$. This is a special case of two-stage sampling that is the same as a single-stage, simple random sample of one branch. We found earlier that the variance among the 135 branches was 1762. The exact variance for a simple random sample of one branch is:

$$(\frac{135 - 1}{135}) \quad \frac{1762}{1} = 1748.8$$

4.5 UNEQUAL PROBABILITY OF SELECTION AT BOTH STAGES

As a further exposition of the theory for two-stage sampling, suppose a sample of trees is selected with replacement and with selection probabilities proportional to trunk size. Also, suppose that the method of sampling at the second stage is the random-path method, RP-PPS, that was discussed in Chapter III. You may recall that the random-path method was presented in the context of sampling with replacement.

When the sampling at both stages is with unequal probability, the estimator of the population total Y is:

$$\hat{y}_t = \frac{\sum_i^m \sum_j^{n_i} \left[\frac{y_{ij}}{p_{ij}}\right]}{n} \tag{4.18}$$

where $$n = \sum_i^m n_i$$

$$p_{ij} = p_i p(j|i)$$

p_i is the selection probability for the i^{th} psu in the sample, and

$p(j|i)$ is the selection probability for the j^{th} ssu given that its psu has been selected.

Consider the quantity $\frac{y_{ij}}{p_{ij}}$ in the estimator. When the value for a unit in the sample (in this case, y_{ij}) is divided by its selection probability (in this case, p_{ij}) the quotient is an estimate of the population total. Therefore, $\hat{y}_t$ in Eq. 4.18 is an average of n estimates of Y, one estimate from each branch in the sample.

The subscript "t" was added to $\hat{y}$ because it is an estimator of Y rather than $\bar{\bar{Y}}$. Notice that the estimator does not contain N. In practice, one finds many populations to be sampled where N is unknown. An estimate, $\hat{N}$, of N might be made from a sample and, if needed, $\frac{\hat{Y}}{\hat{N}}$ could be used as an estimate of $\bar{\bar{Y}}$. An estimator of N is obtained by substituting "1" for y_{ij} in 4.18.

Exercise 4.18 *Suppose, for* m=3, *and* n_i=2, *that application of the above method to the apple tree population gives the following sample:*

Population index values of i and j		*Sample index values of i and j*				
Tree	*Path*	*Tree*	*Path*	p_i	$p(j\|i)$	y_{ij}
1	2-2-3	3	1	0.07035	.15996	59.5
1	4-1	3	2	0.07035	.07779	7.0
3	1-2-1-2	1	1	0.2312	.04864	139.3
3	2-3	1	2	0.2312	.05757	125.0
3	3-1-2	2	1	0.2312	.02081	31.3
3	1-2-1-2	2	2	0.2312	.04864	139.3

Tree No. 3 and path 1-2-1-2 *were selected twice. The selection probabilities* P_i *were proportional to* X_i*, the trunk sizes which are presented in Table 4.3. Verify the values of* p_i*. For tree No. 3 in the population, the conditional probabilities,* $P(j|i)$*, are the probabilities in Column* P_4 *of Table 3.2. Thus, the above values of* $p(j|i)$ *and* y_{ij} *for the branches in the sample from this tree were taken from Table 3.2. The values of* $p(j|i)$ *and* y_{ij} *for tree No. 1 are from records not reproduced herein. Using Eq. 4.18 as the estimator, calculate the estimate of the total number of apples. The answer is* 7873*, which is an estimate of* 7199*, the total number of apples including "path" apples (See Table 4.3).*

To find the variance of $\hat{y}_t$, refer to Eq. 4.13 and make two modifications:

(1) for the first stage we want $N^2\sigma_1^2$ instead of σ_1^2 because $\hat{y}_t$ is an estimator of Y rather than $\bar{\bar{Y}}$, and

(2) for the second stage, the part of the formula representing the variance of an estimate of Y_i for a simple random sample of n_i needs to be changed. That is, $(1 - f_{2i}) \frac{N_i^2 S_{2i}^2}{n_i}$ needs to be replaced by the corresponding variance for sampling within the i^{th} psu with pps. Also, $\frac{1}{N^2}$ needs to be dropped. This gives:

$$V(\hat{y}_t) = \frac{1}{m}\left[\sum_i^M P_i\left(\frac{Y_i}{P_i} - Y\right)^2 + \sum_i^M \left(\frac{1}{P_i}\right) \frac{S_{ri}^2}{n_i}\right] \qquad (4.19)$$

where $$S_{ri}^2 = \sum_{j}^{N_i} P(j|i)\left(\frac{Y_{ij}}{P(j|i)} - Y_i\right)^2$$

The subscript r signifies random path.

For Tree No. 3 the values of $P(j|i)$ and the values of $\frac{Y_{ij}}{P(j|i)}$ are in columns P_4 and $\hat{Y}_4$ of Table 3.2. From these two columns the value of S_{ri}^2 for tree number 3 can be computed. The answer, 800,000 is recorded along with other values of S_{ri}^2 in the last column of Table 4.3.

Exercise 4.19 When $n_i=\bar{n}$, *the second term in* [] *of Eq. 4.19 becomes* $\frac{1}{\bar{n}} \sum_{i}^{M} \frac{S_{ri}^2}{P_i}$. *In the problem under consideration,* $P_i=\frac{X_i}{X}$ *where* X_i *is trunk size. From the data in Table 4.3, find the value of* $\Sigma\frac{S_{ri}^2}{P_i}$ *and of* $\Sigma P_i\left(\frac{Y_i}{P_i} - Y\right)^2$. *When your results are substituted in 4.19, you should have:*

$$V(\hat{y}_t) = \frac{1}{m}\left[5{,}322{,}000 + \frac{11{,}462{,}000}{\bar{n}}\right]$$

Exercise 4.20 From the sample data given in Exercise 4.16 estimate the total number of terminal branches on the six trees. Ans. 122.0.

When the equation in Exercise 4.19 is divided by N^2 we obtain:

$$V(\hat{y}) = \frac{1}{m}\left[292 + \frac{629}{\bar{n}}\right]$$

To summarize, the following variance equations have been obtained for three alternative two-stage plans for sampling the small population of apple trees:

$$(1) \qquad V(\hat{y}) = \frac{1}{m}\left[917.1 + \frac{(1-f_2)\ 1367}{\bar{n}}\right]$$

for simple random sampling at both stages, where $\frac{n_i}{N_i}$ is constant and equal to f_2 and $1-f_1$ was assumed to be equal to 1,

$$(2) \qquad V(\hat{y}) = \frac{1}{m}\left[439.7 + \frac{1367 - \bar{n}(58.0)}{\bar{n}}\right]$$

for sampling trees with probability proportional to N_i and a simple random sample of $\bar{n}$ branches from each selected tree, and

$$(3) \qquad V(\hat{y}) = \frac{1}{m}\left[292 + \frac{629}{\bar{n}}\right]$$

for sampling trees with probability proportional to X_i (trunk size) and application of the RP-PPS method to the sample trees.

The results are too limited to provide a basis for generalization.

Table 4.1 Representation of Population Data for Two Stage Sampling[1/]

psu	ssu: 1 ... j ... N_i	psu total	psu mean	Within psu variances
1	$Y_{11} \cdots Y_{1j} \cdots Y_{1N_1}$	Y_1	$\bar{\bar{Y}}_1$	$S_{21}^2 = \frac{\sum_j (Y_{1j} - \bar{\bar{Y}}_1)^2}{N_1 - 1}$
.				
.				
i	$Y_{i1} \cdots Y_{ij} \cdots Y_{iN_i}$	Y_i	$\bar{\bar{Y}}_i$	$S_{2i}^2 = \frac{\sum_j (Y_{ij} - \bar{\bar{Y}}_i)^2}{N_i - 1}$
.				
.				
M	$Y_{M1} \cdots Y_{Mj} \cdots Y_{MN_M}$	Y_M	$\bar{\bar{Y}}_M$	$S_{2M}^2 = \frac{\sum_j (Y_{Mj} - \bar{\bar{Y}}_M)^2}{N_M - 1}$

[1/] A single bar "-" is used for an average of psu totals and a double bar "=" indicates an average of secondary units. A subscript 1 or 2 affixed to S^2 indicates first or second stage variance. See definitions below.

Y_{ij} is the value of the characteristic Y for the j^{th} ssu in the i^{th} psu,

$Y_i = \sum_j^{N_i} Y_{ij}$ is the total of Y for the i^{th} psu,

$Y = \sum_i^M \sum_j^{N_i} Y_{ij} = \sum_i^M Y_i$ is the total of Y for the population,

M is the number of psu's in the population,

N_i is the population number of ssu's in the i^{th} psu,

$N = \sum_i^M N_i$ is the number of ssu's in the population,

$\bar{Y} = \frac{Y}{M}$ is the population mean per psu,

$\bar{\bar{Y}} = \frac{Y}{N}$ is the population mean per ssu,

$\bar{\bar{Y}}_i = \frac{Y_i}{N_i}$ is the average value of Y per ssu in the i^{th} psu,

$\bar{N} = \frac{N}{M}$ is the average number of ssu's per psu,

$S_{2i}^2 = \sum_j^{N_i} \frac{(Y_{ij} - \bar{\bar{Y}}_i)^2}{N_i - 1}$ is the variance among ssu's in the i^{th} psu, and

$S_1^2 = (\frac{1}{\bar{N}^2}) \frac{\sum_i^M (Y_i - \bar{Y})^2}{M-1}$ is the variance among psu's on the basis of one ssu.

Table 4.2 Components of Variation for a Hypothetical Population

psu	1	2	3	4	5	Y_i	$\bar{\bar{Y}}_i$	S^2_{2i}
	Values of Y_{ij}							
1	67	45	51	20	35	218	43.6	308.8
2	32	27	82	39	18	198	39.6	620.3
3	14	25	21	30	28	118	23.6	40.3
4	55	48	72	63	88	326	65.2	242.7
						$Y = 860$	$\bar{\bar{y}} = 43.0$	
	Values of $(Y_{ij} - \bar{\bar{Y}})$							
1	24	2	8	-23	-8			
2	-11	-16	39	-4	-25			
3	-29	-18	-22	-13	-15			
4	12	5	29	20	45			
	Values of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$							
1	0.6	0.6	0.6	0.6	0.6			
2	-3.4	-3.4	-3.4	-3.4	-3.4			
3	-19.4	-19.4	-19.4	-19.4	-19.4			
4	22.2	22.2	22.2	22.2	22.2			
	Values of $(Y_{ij} - \bar{\bar{Y}}_i)$							
1	23.4	1.4	7.4	-23.6	-8.6			
2	-7.6	-12.6	42.4	-0.6	-21.6			
3	-9.6	1.4	-2.6	6.4	4.4			
4	-10.2	-17.2	6.8	-2.2	22.8			

$S^2 = 487.053$ is the variance among the 20 values of $(Y_{ij} - \bar{\bar{Y}})$

$S_1^2 = 293.707$ is the variance among the 4 values of $(\bar{\bar{Y}}_i - \bar{\bar{Y}})$

$S_2^2 = 303.025$ is the average of the variances of $(Y_{ij} - \bar{\bar{Y}}_i)$ within psu's. Within the first psu the variance is:

$$\frac{23.4^2 + 1.4^2 + 7.4^2 + (-23.6)^2 + (-8.6)^2}{4} = 308.8.$$

Table 4.3 Summary Data for Six Apple Trees [1/]

Tree	No. of Terminal Branches N_i	No. of Apples on Terminal Branches Y_i	Within Tree Variance DS-EP S^2_{2i}	Trunk Size in Sq. In. X_i	Total No. of Apples on Tree Y'_i	Within Tree Variance RP-PPS S^2_{ri}
1	13	213	259	7.0	214	22,000
2	27	1,388	1,147	20.0	1,448	478,000
3	26	1,850	2,184	23.0	1,901	800,000
4	20	1,592	3,106	16.5	1,658	350,000
5	19	402	241	13.5	403	79,000
6	30	1,528	892	19.5	1,575	513,000
Total	135	6,973		99.5	7,199	

[1/] The values of N_i, Y_i, and S^2_{2i} are from Table 2.6. The values of Y_i and S^2_{2i} are labeled Y_h and S^2_{Yh} in Table 2.6. "Path apples" are not included in Y_i and S^2_{2i}. The values of Y'_i and S^2_{ri} include the path apples and are taken from Table 3.3. The subscript "r" refers to random path.

DS-EP and RP-PPS refer to the method of sampling a tree as discussed in Chapter III.

Table 4.4 Variances for Alternative Sample Allocations

	m	$\bar{n}$	$V(\hat{y})$	Components: Among psu's [1/]	Components: Within psu's
(1)	1	4	1225.8	917.1	308.7
(2)	2	2	797.6	458.6	339.0
(3)	4	1	583.5	229.3	354.2

[1/] Assumes f_1 is negligible.

☆ U. S. GOVERNMENT PRINTING OFFICE : 1978 261-494/151